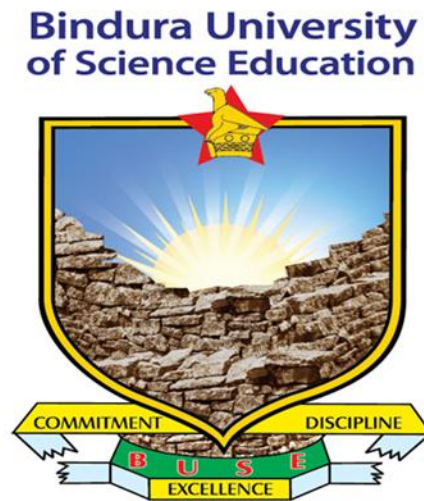


BINDURA UNIVERSITY OF SCIENCE EDUCATION
FACULTY OF SCIENCE AND ENGINEERING
DEPARTMENT OF STATISTICS AND MATHEMATICS



TIME SERIES ANALYSIS OF CHOLERA CASES IN BINDURA DISTRICT (2014-2023)
BY

MADELINE MAGARIRO

A DISSERTATION SUBMITTED IN PARTIAL FUFILMENT OF THE
REQUIREMENTS
FOR BSC.HONOURS IN STATISTICS AND FINANCIAL MATHEMATICS
SUPERVISOR: DR. T.W. MAPUWEI

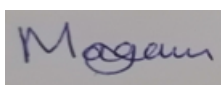
2024

APPROVAL FORM

The undersigned certify that they have supervised, read and recommend to the Bindura University of Science Education for the acceptance of a research dissertation entitled:

Time series analysis of cholera cases in Bindura District in the fulfilment of the requirements for the Bachelor of Science Honors Degree in Statistics and Financial Mathematics.

MADELINE MAGARIRO



(B202554B)

(signature)

DR. T.W. MAPUWEI



(supervisor)

(signature)

DR. MAGODORA



(chairperson)

(signature)

DEDICATION

This effort honours God who has been my pillar of strength throughout the program. I gratefully dedicate this research project to my lovely family for their encouragement and financial support which empowered me to undertake and complete this work.

ACKNOWLEDGEMENTS

First and foremost, I give glory to the Almighty God, the source of wisdom and knowledge, for providing the inspiration, and means to undertake this research. I am grateful for his blessings and guidance throughout this dissertation. I would like to express my gratitude to my supervisor, Dr. Mapuwei for his encouragement, patience, invaluable guidance and constructive feedback throughout this duration. This shaped the direction and quality of this work. I would like to acknowledge the love and financial support of my family and the encouragement of my friends who have been a constant force of motivation and inspiration throughout this journey. My appreciation goes to my colleagues who helped during the time of studying.

ABSTRACT

Cholera outbreaks raise a significant public health challenge in many developing countries, including Zimbabwe. Bindura District which is located in the province of Mashonaland Central in Zimbabwe have experienced severe outbreaks over the past decade. The aim of the study is to investigate the trends, seasonality of the cholera cases and cholera deaths in Bindura District from 2014 to 2023 and forecast the future cholera outbreaks. A quantitative research design was employed to analyze 40 quarterly observations of cholera cases and deaths, spanning a period of ten years. The data was obtained from the Ministry of Health and Child Care's Bindura District office and the analysis was conducted using R Studio software. Initial analysis revealed that the data exhibited non-stationarity and autocorrelation. To address these issues and develop an effective model, auto ARIMA algorithm was employed to identify the optimal ARIMA specification, which accounted for the non-stationarity and autocorrelation in the data. The Akaike Information Criterion (AIC) was employed to select the optimal models for cholera cases and deaths, respectively. The optimal models were identified as ARIMA (2,1,0) (0,1,1) [4] for cholera cases and ARIMA (0,0,1) (0,1,0) [4] for deaths. The data was used for forecasting and the results showed that both cholera cases and deaths were expected to fluctuate until 2026. Mean squared error, mean absolute error and root mean square error metrics were utilized to analyze prediction accuracy of the model. The performance metrics were low indicating that the ARIMA models had good predictive accuracy and it demonstrates the effectiveness of ARIMA model in capturing the patterns and trends in the data. The forecasting results revealed a discordant trend between cholera cases and deaths. While cholera cases are predicted to increase, cholera deaths are not expected to follow a similar pattern, suggesting a decoupling of the cholera cases and deaths. Since the cholera cases and deaths are fluctuating, there is need to strengthen disease surveillance systems to detect cholera outbreaks early, increase public health measures during peak seasons and conduct regular public awareness campaigns to educate communities about cholera prevention, symptoms and treatment.

TABLE OF CONTENTS

TITLE:	i
APPROVAL FORM	ii
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	v
LIST OF FIGURES	ix
LIST OF TABLES	x
LIST OF ABBREVIATIONS	xi
CHAPTER 1	1
INTRODUCTION	1
1.0: Introduction	1
1.1: Background	1
1.2: Problem Statement	3
1.3: Aim of The Research	3
1.4: Research Objectives	3
1.5: Research Questions	4
1.6: Scope of the Study	4
1.7: Significance of the Study	4
1.8: Assumptions.....	4
1.9: Limitations	5
1.10: Definition Of Terms	5
1.11 Conclusion.....	5
CHAPTER 2: LITERATURE REVIEW	6
2.0: Introduction	6
2.1: Theoretical Literature.....	6
2.1.1 Components of Time Series	6
2.1.2 Assumptions of time series	7
2.1.3 Models in Time Series.....	8
2.2 Empirical Literature	9
2.4: Research Gap	12
2.5: Conceptual Framework	13

CHAPTER 3.....	14
RESEARCH METHODOLOGY	14
3.0: Introduction	14
3.1: Research Design.....	14
3.2: Data Sources.....	14
3.3: Target Population	14
3.4: Research Instrument.....	15
3.5: Data Collection.....	15
3.6: Description of Variables and Expectation.....	15
3.7: Data Analysis	16
3.7.1 Diagnostic Tests.....	16
3.7.2 Analytical Model	17
3.7.3 Model Validation.....	17
3.7.3.1 Mean Squared Error	17
3.7.3.2 Mean Absolute Error	18
3.7.3.4 Mean Absolute Percentage Error	18
3.8: Ethical Considerations	18
3.9: Conclusion.....	19
CHAPTER 4: DATA PRESENTATION, ANALYSIS AND INTERPRETATION	20
4.0: Introduction	20
4.1: Descriptive Statistics	20
4.2 Pre – Tests/ Diagnostic Tests	21
4.2.1 Normality test	21
4.2.2 Stationarity test	23
4.2.3 Autocorrelation test	23
4.2.4 Seasonal trends	24
4.3: Model Output	25
4.3.1 Residuals	27
4.3.2 Forecasting.....	30
4.4: Model Validation	31
4.5: Conclusion.....	33
CHAPTER 5: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	34
5.0 Introduction	34

5.1 Summary of Findings	34
5.2 Conclusions	34
5.3 Recommendations	35
5.4: Areas of Further Study	35
5.5: Chapter Summary.....	35
REFERENCES.....	36
APPENDICES	38

LIST OF FIGURES

Figure 1: Components time series.....	7
Figure 2: Proposed conceptual model.....	13
Figure 3: Boxplot for cholera cases	21
Figure 4: Boxplot for cholera deaths	21
Figure 5: Histogram for cholera cases	22
Figure 6: Histogram for cholera deaths	22
Figure 7: ACF and PACF plots for cholera cases	23
Figure 8: ACF and PACF plots for cholera deaths	24
Figure 9: Seasonal trend for cholera cases.....	24
Figure 10: Seasonal trend for cholera deaths.....	25
Figure 11: ACF and PACF plots of cholera cases residuals	27
Figure 12: ACF and PACF of cholera deaths residuals	28
Figure 13: Patterns for cholera cases residuals	28
Figure 14: Patterns cholera deaths residuals	29
Figure 15: Seasonal pattern for predicted cholera cases	31
Figure 16: Seasonal pattern for predicted cholera deaths	31

LIST OF TABLES

Table 1: Descriptive statistics	20
Table 2: Summary for ARIMA (2,1,0) (0,1,1) [4] model.....	26
Table 3: Summary for ARIMA (0,0,1) (0,1,0) [4] model.....	26
Table 4: Predicted cholera cases	30
Table 5: Predicted cholera deaths	30

LIST OF ABBREVIATIONS

ACF - Autocorrelation Function

ADF - Augmented Dickey-Fuller

AIC - Akaike Criteria

AR - Autoregressive

ARIMA - Autoregressive Moving Average

GIS - Geographic Information System

ICDDR, B - International Centre for Diarrheal Disease Research, Bangladesh

MA - Moving Average

MAE - Mean Absolute Error

MAPE - Mean Absolute Percentage Error

MoHCC - Ministry of Health and Child Care

MSE - Mean Square Error

PACF - Partial Autocorrelation Function

RMSE – Root Mean Square Error

SI - Seasonal Difference

UNICEF - United Nations Children's Funds

WHO - World Health Organization

CHAPTER 1

INTRODUCTION

1.0: Introduction

A serious public health issue that impacts measures of cholera have been conducted. However, there remains a knowledge gap where millions of people are affected by cholera. Cholera is an acute diarrheal infection caused by taking in of food or water which has the bacterium *Vibrio cholerae* (WHO, 2023). Despite that there is an improvement on the water treatment and sanitation, cholera outbreaks remain a threat to the public health causing untold lives and taking a lot of lives. Numerous researches on the causes, consequences, and preventative were done regarding the evolving patterns and trends of the disease over time. This chapter will highlight the study's background, problem statement, research objectives, and questions in order to close this gap. This chapter also covers the study's limitations, significance, and scope.

1.1: Background

The disease known as cholera began in the reservoir in the Ganges delta of India and spread throughout the 19th century (Lopez et al, 2015). Millions of people perished in more pandemics that struck every continent. Beginning in South Asia 1961, the seventh pandemic spread to Africa in 1971 and America in 1991. There is a high number of reported cases of Cholera to World Health Organization for the past few years. Cholera is a major public health issue which has affected the global with about 1.3 to 4 million cases and 21 000 to 143 000 deaths which occur each and every year (World Health Organization, 2020). It has been discovered that in most cases, people living in poverty, inadequate access to safe water and basic sanitation are the ones affected by cholera the most (Semenza, 2017).

Conflict, unplanned urbanization and climate change intensify the risk of cholera outbreaks (Kouadio, 2018). Oral rehydration is good to treat those with mild symptoms of cholera disease (Ali, 2019). However, severe instances need to be treated right away. with intravenous fluids and antibiotics (World Health Organization, 2019). Prevention and control of cholera transmission rely on the provision of safe water and basic sanitation, as well as good hygiene practices. Effective interventions, such as oral cholera vaccines, have also been shown to reduce the incidence of

cholera (Lopez, 2018). Furthermore, studies have highlighted the importance of addressing the root causes of cholera, including poverty, lack of access to healthcare, and poor water and sanitation infrastructure (Semenza, 2017).

Leg cramps, vomiting, and frequent water diarrhea are the symptoms of a severe infection, which affects 20% of cases. Mild infections typically have no symptoms at all. Poor neighborhoods with dense populations and inadequate access to clean water and sanitary facilities are susceptible to cholera. According to (WHO & UNICEF, 2019), over 2 billion people drink water contaminated with stool, and 663 million people worldwide lack access to safe drinking water.

Ali et al., (2015) reported that 2.9 million cases and 95,000 deaths of cholera are reported annually in 69 endemic countries. When three years have passed and there is proof of local transmission, an area is considered endemic for cholera (WHO, 2022). Zimbabwe is classified as an endemic nation. The ongoing outbreaks in Zimbabwe are brought on by a lack of access to basic health care, clean water, and sanitation. The most severe cholera outbreak, which resulted in the nation's crumbling economic structure, occurred between 2008 and 2009. August 2008 saw the beginning of the outbreak in Chitungwiza, a neighborhood near Harare. 4288 fatalities and 98592 cases overall had been identified by the end of April 2009, at which point 95% of the nation's districts had been impacted by the epidemic. 4.3% was the case fatality rate. UNICEF gave Zimbabwe a \$5 million boost to combat the cholera outbreak. It also started to provide water treatment chemicals, as well as medications for affected individuals and water supplies, to local authorities so that residents could have access to safe drinking water.

Another severe cholera outbreak occurred in Zimbabwe; between September 6, 2018, and March 26, 2019, there were 371 confirmed cases and 68 deaths reported. On October 3, 2018, a two-dose oral cholera vaccination campaign was launched in Harare. The second round took place in 2019 in March and April. Zimbabwe is one of the confirmed 24 countries that have experienced cholera outbreaks since the start of 2023. As of November 20th, Zimbabwe's Ministry of Health and Child Care recorded 8,230 suspected cases of cholera, 1,301 laboratory confirmed cases (9,531 total cases), and 209 fatalities. Significant health has been raised by the outbreak as of 29 February 2023 especially in Mashonaland Central province which recorded over 2850 cases 94 deaths accounting for 19.5% of the national burden. Bindura district in Mashonaland Central province is one of the places which has been affected by cholera the most and there are several cases which has been

reported (World Health Organization, 2023). The flourishing mining industry in Bindura a temporary population of artisanal miners, who access water from contaminated sources like abandoned open cast mines and the Pote river, creating a breeding ground for cholera (Mudyiradima, 2023). Lack of access to proper sanitation have worsened the situation hence increasing the spread of the disease (UNICEF, 2023). In response, a treatment center for cholera has been placed at Tafuna shopping center to donate and provide medical care to the affected individuals (MoHCC, 2023).

1.2: Problem Statement

Cholera outbreaks has been a heavy burden to the public health of Zimbabwe causing high mortality rates, widespread morbidity and significant economic costs. The country's cholera control and prevention initiatives are hindered by the surveillance system that struggles to provide timely and accurate information, limiting the effectiveness forecasting and response strategies. The current system, which relies on passive reporting which leads to incomplete and delayed reporting causing delayed responses, whereas the rapidly changing of cholera's outbreaks nature, influenced by factors like weather conditions, seasonality and human behavior needs a more dynamic and proactive approach to surveillance and response.

1.3: Aim of The Research

The main aim of this study is to identify patterns, trends and to forecast future cholera cases over time in Bindura District, assisting to understand the dynamics of the disease.

1.4: Research Objectives

1. Utilize quarterly data to identify long term trends in reported cholera cases and cholera related deaths over the ten-year period in Bindura District.
2. Investigate seasonal patterns in cholera outbreaks to understand if there are specific quarters or times of the year when reported cases and deaths stands to peak.
3. Develop ARIMA model to forecast future trends in reported cases and deaths in Bindura District.

1.5: Research Questions

1. What are the long-term trends in reported cholera cases and deaths over a 10-year period?
2. Which quarters or months typically have the most or least recorded cases and fatalities of cholera.
3. Can ARIMA model be developed accurately to forecast future trends in the quarterly reported cases and deaths?

1.6: Scope of the Study

The goal of the study is to examine a time series analysis of cholera cases and deaths in Bindura District, Zimbabwe. The data spans from 2014 to 2023 and the data will be a quarterly data of cholera cases and deaths from each year. The data include the infected people in the urban and rural areas of the Bindura District.

1.7: Significance of the Study

This study is significant because it improves public health surveillance an effective time series model for accurate cholera cases and deaths forecasting. By better understanding the temporal dynamics of cholera in Zimbabwe, the study will aid public health officials in focusing on interventions more effectively for optimizing cholera control and prevention. Additionally, by identifying the patterns and trends in cholera cases and deaths, the study can help the public health professionals anticipate and prepare for outbreaks using data from past outbreaks, potentially reducing the number of cases and deaths.

1.8: Assumptions

1. The data is independent (the number of cases in one quarter is not related to the number of cases in another quarter).
2. The time series data has a constant mean and variance over time hence it shows stationarity.
3. The number of instances is distributed uniformly over all months.

4. The chosen time series model is good for the data and can accurately capture the trends and future forecast.

1.9: Limitations

This study has various restrictions and it is probable that some cholera cases go undetected because the statistics on the disease is dependent on reported cases, therefore some may not report to the clinics to be recorded. Also, there might be errors in the data entry. Furthermore, the study ignores variables that may also have an impact on the prevalence of cholera, such as population density, water quality, and sanitation. The study is more focused on the existing data which may have limitations on the accuracy and completeness.

1.10: Definition Of Terms

Public health surveillance

The systematic, continuous collection, analysis and interpretation of health-related data needed for the planning, implementation and evaluation of public health practices (WHO, 2018).

Time Series

It is a set of data points that are measured at regular intervals and are commonly used to forecast and analyze trends and patterns in a variety of fields (Reinsel, 2016).

Forecasting

The process of projecting future values or results using statistical models and historical data (Armstrong, 20216).

1.11 Conclusion

In conclusion, this chapter has outlined the particular goals and objectives of the study as well as its background and significance. Assumptions, constraints, and the study's scope had all been covered. It is envisaged that a deeper comprehension of cholera outbreaks and the development of successful interventions to lessen the disease's impact on the populace will be possible through the application of time series analysis.

CHAPTER 2: LITERATURE REVIEW

2.0: Introduction

The literature review critically examines current research on cholera epidemiology, transmission dynamics and use of time series analysis in public health surveillance. This chapter is summarizing the state of knowledge currently known about these subjects by highlighting significant discoveries, methodological strategies and gaps in the body literature. It is also reviewing and summarizing the works of the past researchers on the issue of cholera.

2.1: Theoretical Literature

According to Ayling et al. (2023), time series analysis is a powerful tool for understanding the patterns of cholera outbreaks overtime. The time series approach can help identify the cyclical patterns of outbreaks and the factors that influence theses cycles. There are several theories which talk about the time series analysis used on cholera.

2.1.1 Components of Time Series

Trend

Trend theory, in the context of time series analysis, refers to a long-term pattern or direction in a time series that persists over time often in an upward or downward slope or a cyclical pattern that repeats over time (Brockwell & Davis, 2016). There are several types of trends, including upward, downward, horizontal, and curvilinear (Makridakis et al. 2018). This trend is important because it helps to understand the latent patterns and directions in a time series which can inform forecasting, decision making and detection (Bell et al. 2014).

Seasonality

Seasonality refers to the periodic and predictable fluctuations in a time series that occur at fixed intervals such as weekly, monthly or yearly cycles (Hyndman et al, 2014). Seasonality theory suggests that weather and climate patterns can play a significant role in the spreading of infectious diseases like cholera. Cholera cases increase more when the temperatures are high, therefore countries like Zimbabwe where the majority of cholera cases occur during the hot, rainy season from October to March, the warm temperatures and frequent rainfall create conditions that increase the spreading of cholera. Understanding these patterns can help public health officials prepare and

implement preventive measures, such as improving water treatment and sanitation facilities, to reduce the spread of cholera during these high-risk periods.

Cyclical

Is a type of pattern or component that exhibits periodic fluctuations, typically with a fixed frequency period (Davis et al., 2016).

Irregularity

Irregularities are anomalies or unforeseen occurrences that deviate from the general pattern or trend. These phenomena are frequently unanticipated and cannot be explained by traditional seasonal or cyclical components (Reinsel et al., 2016).

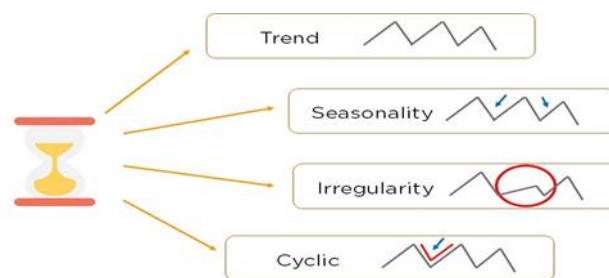


Figure 1: Components time series

2.1.2 Assumptions of time series

Stationarity

A time series is termed stationary if its statistical features such as mean, variance and covariance are constant over time (Hyndman and Athanasopoulos, 2018).

Normality

Normality implies that the data has a normal distribution, often known as bell curve. A normal distribution is symmetrical around mean, with no skewness and outliers (Hyndman and Athanasopoulos, 2018).

Independence

Refers to the assumption that every observation in the series is independent and uncorrelated with others (Hyndman and Athanasopoulos, 2018).

2.1.3 Models in Time Series

Autoregressive Model

One particular kind of regression model in which the dependent variable is based on past values is the autoregressive model (AR). This suggests a relationship between the forecast and previously observed values. The idea of partial autocorrelation is the basis of this approach. An AR model of order p is represented by the equation below:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t \dots\dots\dots (2.1)$$

Moving Average (MA)

It does not have any seasonal elements that are somewhat random (Pannerselvam, 2005). When there is a slight trend or season, moving average models have the benefit of being able to predict stocks or goods with consistent demand. They are also helpful in determining locations of support and resistance. Conversely, seasonal impacts and cycles, for example, have no effect on MA because they are caused by other factors. The formula for figuring out the Moving Average is as follows:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}, \dots\dots\dots (2.2)$$

Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model is a well-liked statistical analysis model for time series data. It combines the concepts of autoregression, differencing, and moving averages to estimate future values and capture the complex dynamics of time series data. Three elements comprise ARIMA: autoregressive (AR), integrated (I), and moving average (MA). Using historical data, the ARIMA

model is used to anticipate values for the future, identify patterns, and analyze time series data. The formula for ARIMA is as follows:

$$y_t = c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t \dots \dots \dots (2.3)$$

Seasonal Autoregressive Integrated Moving Average (SARIMA)

By adding seasonal components, the SARIMA model is a time series forecasting technique that expands upon ARIMA. For modelling and forecasting time series data, it is an effective technique. A SARIMA model would normally be implemented by determining the seasonal components, the AR, and the MA, and then fitting the model to the data to produce forecasts. The SARIMA equation is displayed below.

$$(1 - \phi_1 B) (1 - \phi_1 B^4) (1 - B) (1 - B^4) y_t = (1 + \theta_1 B) (1 + \theta_1 B^4) \epsilon_t, \dots \dots \dots (2.4)$$

B is the Backshift operator

2.2 Empirical Literature

Mashiri et al, (2018) carried out a study looking at the seasonal cholera incidence forecasts in Zimbabwe. The primary objective of this study was to create a dependable and accurate model using ARIMA modelling. They suggested developing a model that would be able to forecast the occurrence of cholera in the future by utilizing this seasonal change. Data was collected from the Ministry of Health and Child Care, which gathers information from clinics, hospitals, and other healthcare facilities and this data was from 1990 to 2012. The model exhibited remarkable accuracy with a mean absolute percentage error (MAPE) of 3.5%, and also had a high degree of accuracy with a mean absolute error (MAE) of 2.4 cases per week. This study's flaw was using outdated data from 1990 to 2012, so the results may not be generalizable to more recent data.

Another study focused on time series analysis approach to investigate the epidemics of cholera that occurred in Zimbabwe between 2009 and 2019. The primary goal was to investigate cholera trends in Zimbabwe, with a particular emphasis on finding the disease's seasonal and trend characteristics (Mangwiro et al, 2020). They wanted to see if there were any differences in cholera trends over time. Another goal was to determine the effect of immunization on cholera incidence.

Data was gathered from Zimbabwe's Ministry of Health and Child Care and the World Health Organization. The Ministry of Health and Child Care provides data on the number of cholera cases reported in Zimbabwe from 2009 to 2019. The World Health Organization provides information on immunization campaigns and better sanitation measures performed in Zimbabwe during this time. The impact was analyzed using data from the Zimbabwe Meteorological Service. The Mann-Kendall test was used to determine whether there was a statistically significant trend in the number of cholera cases over time, autoregressive integrated moving average (ARIMA) modelling to forecast future cholera incidence, and multiple linear regression analysis to assess the impact of various factors on cholera incidence. The study's findings revealed that cholera outbreaks in Zimbabwe typically occurred during the rainy season, and that the incidence of cholera was positively connected with temperature.

Chahwanda et al. (2019) used spatial-temporal analysis to investigate the patterns and dynamics of cholera outbreaks in Zimbabwe. Finding the regions and times of year when cholera outbreaks are most likely to occur was the goal. They also sought to investigate the connections between different environmental and socioeconomic factors and the frequency of cholera outbreaks. The information used in this study was given by Zimbabwe's Ministry of Health and Child Care. The data included the date, location, and number of cases for each outbreak. The study also included data on vegetation cover, temperature, and precipitation as environmental factors. In addition, socioeconomic and demographic data were used, such as the density of health facilities, the percentage of the people living in poverty, and population density. The occurrence of outbreaks was examined using exploratory spatial analysis to look for any spatial trends. Next, they employed statistical modelling to determine the variables linked to the occurrence of cholera outbreaks. These models included generalized linear models, Bayesian models, and space-time models. The authors also used GIS tools to visualize the results of the statistical models. Finally, they used Bayesian model averaging to identify the most important predictors of cholera outbreaks.

Mwayi et al, (2015) analyzed data from cholera outbreaks in Zimbabwe from 2009 to 2012. The theme was to use spatial and temporal data to understand the risk factors for cholera outbreaks in Zimbabwe. The objectives of the study were to determine the risk factors for cholera outbreaks in Zimbabwe, the regional and temporal patterns of cholera incidence, and the distribution of cholera incidence over time and geography. Information was gathered from the World Health

Organization, the Ministry of Health and Child Care, which keeps track of all cholera cases that are reported in the nation, and the Famine Early Warning Systems Network, which offers data on Zimbabwe's rainfall and drought patterns. The data was analyzed using a Bayesian spatiotemporal model in order to determine the risk factors associated with cholera outbreaks. The study's findings demonstrated the high spatial and temporal clustering of cholera epidemics. According to the model, places with a larger population density and rainfall had a higher probability of experiencing cholera epidemics. The model also predicted that there would be more cholera outbreaks during the rainy season. Finally, the study found that there was a strong interaction between space and time, meaning that the risk of a cholera outbreak was not only influenced by the location of the outbreak but also by the time of year. The gap in the study was that the data was limited to the period from 2009 to 2012, so it did not capture any recent changes in cholera incidence. Therefore, my research will extend the period.

Mukandavire et al, (2018) focused on analyzing the patterns of cholera cases over time. Their main focus was to see the temporal patterns in Zimbabwe and identify if there were any seasonal or cyclical trends in cholera incidence. From 2008 to 2015, information on cholera cases was gathered from the Ministry of Health and Child Care in Zimbabwe. Along with the location and date of each case, the data also included information on the number of cases and deaths. The information was gathered from medical facilities around Zimbabwe, and its writers utilized it to make a map of cholera cases in that nation. To find places with inadequate cleanliness, they also analyzed data from satellites. After analyzing the data with the ARIMA model, it was discovered that the number of cholera cases was cyclical, peaking every five to six years. Poisson regression was employed in the study to determine the risk factors for the spread of cholera in Zimbabwe.

Chinake et al, (2018) did a study of assessing temporal patterns of cholera incidence in the basin of Limpopo, which includes the parts of Botswana, Mozambique, and Zimbabwe. The main aim of the study was to identify the spatial distribution of cholera outbreaks in the region and to understand the temporal patterns of cholera incidence, including seasonality and trends over time. It also aimed to identify potential risk factors for cholera outbreaks, such as socioeconomic factors, sanitation, and climate and finally the study aimed to develop a spatiotemporal model that could be used to predict cholera outbreaks in the Limpopo basin. The study used data from the African Water Atlas to create a GIS database of the Limpopo basin. This database included information on

the population, water resources, and climate in the region. The data of the study was taken from the International Centre for Diarrheal Disease Research, Bangladesh (ICDDR, B) to obtain the cholera incidence data for the region. The study also used data from the Zimbabwe Ministry of Health and Child Care to obtain information on health facilities and sanitation. ARIMA model was used with spatial and temporal data to predict future cholera outbreaks in the Limpopo basin. The ARIMA model with spatial and temporal data predicted cholera outbreaks in the Limpopo basin accurately. The model also showed that cholera incidence was highly seasonal and the outbreaks were peak during the rainy season.

One study conducted by Tapera, Dube, and Muchenje (2016) focused on modeling and forecasting cholera cases in Zimbabwe using an ARIMA model. The study used monthly cholera data from 2008 to 2014 and found that the ARIMA model was capable of accurately predicting the number of cholera cases in the country.

2.4: Research Gap

There are several gaps that were found on these studies. The main gap was on the data. Most researchers were collecting data which were covering five years or less. Also, they were not collecting recent data, for an example, the paper could be established in 2018 but they used the data that covered 2008-2009. Also, data was collected from the websites and not collecting it directly from the institution.

2.5: Conceptual Framework

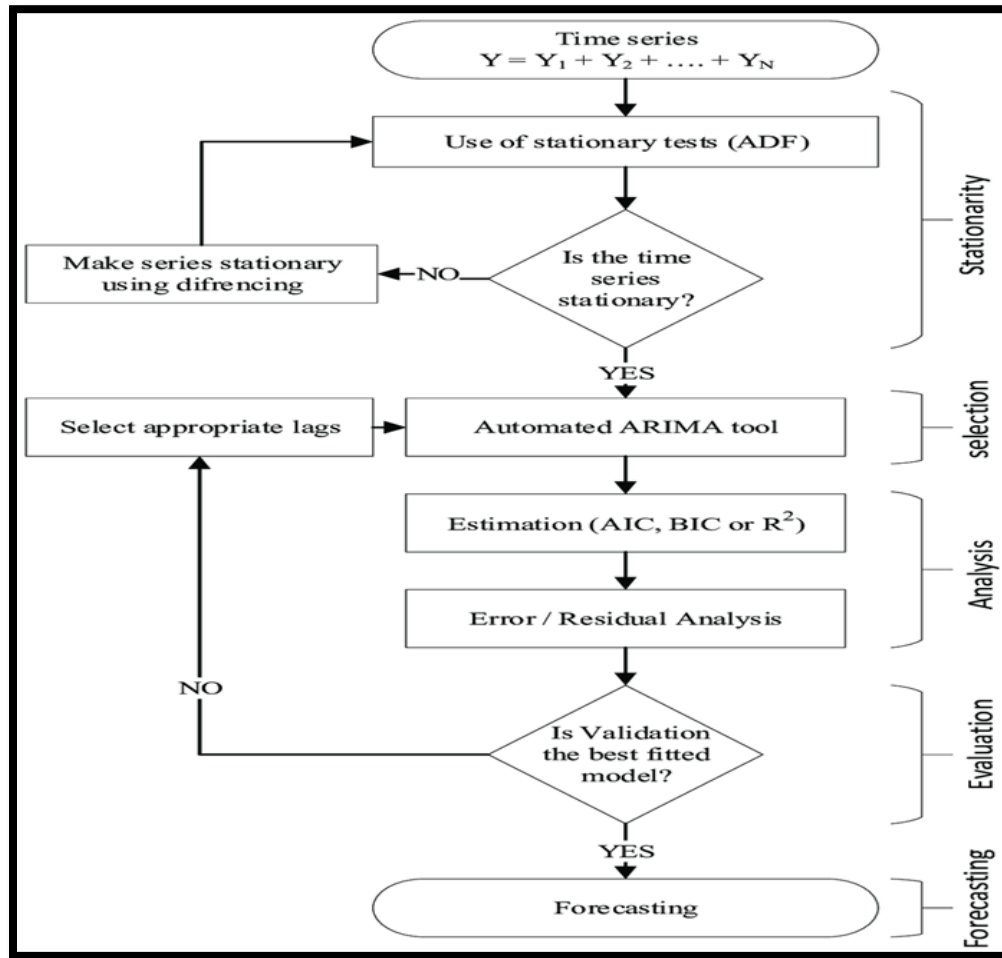


Figure 2: Proposed conceptual model

2.6: Conclusion

In summary, this literature review highlights the importance of time series analysis in understanding the patterns and dynamics of cholera cases in Zimbabwe and it also highlighted the gaps in other studies.

CHAPTER 3

RESEARCH METHODOLOGY

3.0: Introduction

This chapter is explaining the research methodology used by the researcher to investigate time series analysis in forecasting cholera cases in Zimbabwe. In this chapter, there will be a detailed description of the data sources, research design, target population, data collection method, data analysis techniques and ethical considerations that guided the study.

3.1: Research Design

A research design, according to Kothari et al. (2004), is the conceptual framework that the study is carried out inside. A quantitative descriptive design was conducted by the researcher on this study to describe the characteristics of the data and making predictions of future values. This research design was to ensure that this study is valid, relevant and meaningful.

3.2: Data Sources

The data used by the researcher was collected from the Ministry of Health and Child Care Bindura district office. The type of data was secondary data, which was in the institution's electronic software called the Demographic Health Information System 2. The advantage of using the DHIS 2 as the data source is the ability to access high-quality and comprehensive data on a wide range of health indicators. DHIS 2 has a limitation whereby the data may be inaccurate. Secondary data has several advantages, including the fact that the data is readily available, saving time, and providing access to large amounts of data that may not be available on primary data.

3.3: Target Population

The targeted population for this study is all people who were living in urban and rural areas of Bindura District, Zimbabwe who have been diagnosed with cholera and reported at clinics, hospitals and health centers between January 2014 and December 2023. This includes individuals

of all gender and age who have medical attention for cholera related symptoms and have been laboratory confirmed as cholera cases.

3.4: Research Instrument

The researcher used a computer to access the Demographic Health Information System of Ministry of Health for Bindura district to collect data that was suitable for the research. This was the best instrument to use because it is compatible and easy to use.

3.5: Data Collection

The researcher collected secondary data from the Ministry of Health and Child Care in Bindura district. The researcher accessed the database of the DHIS to extract the data. This saved time of the researcher because the data has been already collected and processed.

3.6: Description of Variables and Expectation

The first variable of interest is the number of reported cases, representing the number of reported cholera cases over time. The expectation was to examine whether there is a rise in reported cholera cases that is corresponding to an increase in cholera-related deaths, a decrease, or no significant change. This analysis will provide insights into the severity and impact of cholera outbreaks on mortality.

Time was represented in quarters and the expectation was to use time as the independent variable to assess the temporal patterns of cholera cases and deaths. By analyzing the trends over the 10-year quarterly period, the expectation is to identify any seasonal or long-term patterns in the incidence of cholera and its impact on mortality.

Cholera related deaths is another variable which represents how the number of deaths is affected by cholera cases during each quarter over the 10-year period. It provides crucial information about the severity and impact of cholera outbreaks. The expectation is to understand the relationship between reported cholera cases and cholera-related deaths. Specifically, the analysis aims to determine whether changes in reported cholera cases are expected to be followed by corresponding changes in cholera-related deaths, and the magnitude of this relationship.

3.7: Data Analysis

3.7.1 Diagnostic Tests

These are tests used to check the assumptions of a model and to identify potential problems with the model such as non-normality and autocorrelation (Wooldridge, 2019).

3.7.1.1 Stationarity Test

When doing time series, the data needs to be stationary before it fits into the model. Stationary data has constant mean and constant variance (Hyndman et al, 2018). An Augmented Dickey-Fuller test (ADF) was used to test for stationarity using a package in R which provides the `adf()` function to conduct the test. The significance level of the ADF test is 0.05, therefore if the p-value is less than 0.05 it shows that the data is stationary and if it is above 0.05 it shows that the data is non-stationary (Wagner, 2019). The researcher tested for stationarity for both the cases and deaths.

3.7.1.2 Normality Test

Shapiro-Wilk test was used to examine if the dataset is normal distributed. It is based on the skewness and kurtosis of the data. Moreover, it compares the distribution of sample to the distribution of a normal distribution with the same sample size. If the p-value is above 0.05, it shows that the data is normally distributed and if the p-value is below 0.05, it shows that it is not normally distributed (Hyndman, 2019). This test was used on both the cases and deaths to determine if this data was normal. Visual inspection was also used through the use of histogram graphs as well as box plots to check for normality.

3.7.1.3 Autocorrelation test

Autocorrelation Function (ACF) and Partial Autocorrelation (PACF) plots generated in R were used to examine the autocorrelation structure of the time series data of cholera cases. The ACF plot shows how the time series of cholera cases correlate with lagged versions of itself and the PACF shows how the cholera cases correlate partially with lagged versions of itself. Both ACF and PACF shows the autocorrelation structure of the time series data that is suitable for modelling and forecasting purposes. Ljung-Box was also used for testing autocorrelation. This test is used to examine if there is any autocorrelation in residuals from a model. It also helped the researcher towards a more appropriate model.

3.7.2 Analytical Model

ARIMA was used as the model for data analysis. ARIMA model is a statistical model that uses time series data to forecast future values based on the past behavior. ARIMA models are a combination of autoregressive (AR) and moving average (MA) components (Hyndman et al, 2018). (I) stands for integration and this component removes a certain number of trends in the data such as linear trend or quadratic trend to make the data stationary. The order of the model (p, d, q) determines the number of autoregressive, integrated and moving average terms in the model.

AR forecasts series mainly based solely on the past values in the series called lags.

$$AR \text{ formular: } y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t \dots \dots \dots (3.1)$$

MA forecasts series mainly based on the sorely on the past errors in the series called error lags.

$$MA \text{ formular: } y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} \dots \dots \dots (3.2)$$

$$ARIMA \text{ formular: } y_t = c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t \dots \dots \dots (3.3)$$

Akaike Information Criteria (AIC)

It measures the amount of information the amount of information lost when a model is used to approximate the true model and the goal to select the best model for the dataset (Anderson et al. 2019). It helps to select a model that explains data in a better way while avoiding unnecessary complexity which can lead to overfitting. AIC was used for selecting best fitting ARIMA model from among the set of many models. A model with the lowest value shows that it fits the data in a better way.

3.7.3 Model Validation

Model validation is the process of determining whether a selected statistical model is appropriate or not for the dataset (Hyndman,2018). The researcher used metrics such as mean squared error, mean absolute error and mean absolute percentage error to check for model validation.

3.7.3.1 Mean Squared Error

The Mean Squared Error (MSE) was used to measure the average squared difference between the actual values in the data and the predicted values from the model. It is used to evaluate the accuracy

and validity of the data. If the MSE is greater than 0.5, it shows that the model is not performing well and if the MSE is less than 0.1, it shows that the model is an excellent model fit to the data and is able to forecast future values. If the MSE is from 0.1-0.5, it shows that the model is a good to the data (James et al, 2014).

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n} \dots\dots\dots (3.4)$$

y_i is the actual value.

\hat{y}_i is the corresponding predicted value.

n = the number of observations.

3.7.3.2 Mean Absolute Error

The MAE is the average of the mean absolute deviation between the expected values and the actual values (James et al, 2014). It was used to see the desired level of accuracy of the model. If the MAE is less than 0.1, it shows that the model is performing well and the MAE is greater than 0.5, then the model needs some improvements.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \dots\dots\dots (3.5)$$

3.7.3.4 Mean Absolute Percentage Error

It calculates the average absolute percentage difference between predicted and actual values. A low MAPE that is less than 10% indicates that the model is excellent and the predicted values are very close to actual values. If the value is below 25%, it shows good accuracy and the values are reasonably close (James et al, 2014)).

3.8: Ethical Considerations

The researcher discovered some principles when collecting data and this study was conducted with a careful consideration of ethical principles. For a student to be given data, the Ministry of Health needs letter from the school. The other thing is the information given remains confidential and be for research purposes only. The research was done with transparency and clearly explained the research procedures.

3.9: Conclusion

This chapter helps to understand the importance of model validation and diagnostic testing. It also helps to learn how to evaluate the performance of a model on a test dataset. This chapter helped the researcher develop the model that can be applied in practice.

CHAPTER 4: DATA PRESENTATION, ANALYSIS AND INTERPRETATION

4.0: Introduction

This chapter examines the findings from the time series analysis of cholera cases and deaths in Bindura District over the 10-year period. The analysis aimed to understand the long-term trends, seasonal patterns and forecasting of cholera in the district. The chapter presents descriptive statistics of the data, diagnostic tests and it discusses the results of the ARIMA model.

4.1: Descriptive Statistics

Table 1: Descriptive statistics

```
> summary(Cholera)
```

Year	Cases	Deaths
Length:40	Min. : 80.00	Min. :12.0
Class :character	1st Qu.: 97.25	1st Qu.:18.0
Mode :character	Median :110.00	Median :21.0
	Mean :110.00	Mean :21.2
	3rd Qu.:121.25	3rd Qu.:24.0
	Max. :145.00	Max. :30.0

The value of 110 represented the average number of cholera cases reported from 2014-2023. Since the mean and the median were equal, it showed that the distribution of cholera cases is relatively symmetrical and not heavily skewed. It also indicated the normal distribution of data, and the mean and median are both reliable indicators of central tendency. The value of 21.2 (mean) and 21.0 (median) represented the average number of cholera-related deaths reported at the hospital over the time period considered. The mean and the median of cholera related deaths are very close, indicating that the distribution of cholera-related deaths is also relatively symmetrical. Therefore, the data distribution for both cases and deaths appears to be relatively symmetrical, indicating a stable pattern over time.

4.2 Pre – Tests/ Diagnostic Tests

4.2.1 Normality test

The Boxplots were used to assess the normality of the data. As shown by the diagrams below, the box plots for cholera cases and related deaths indicate normality. The box plots also provide a clear presentation of data's distribution which shows that the data meets the assumption of normality.

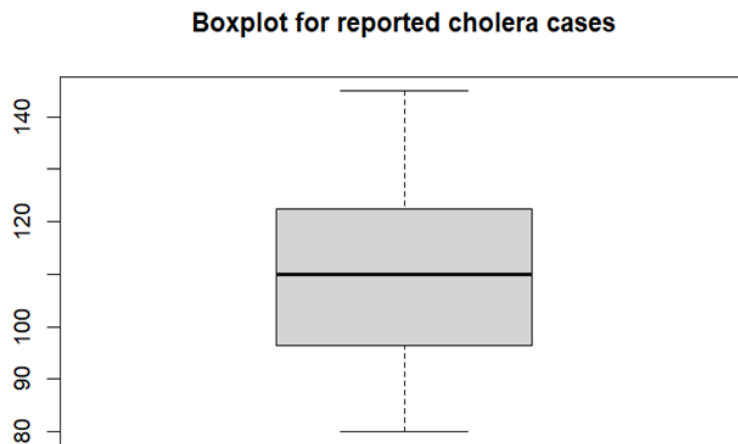


Figure 3: Boxplot for cholera cases

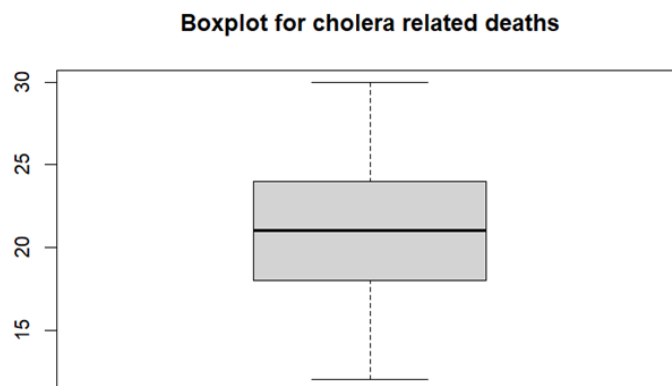


Figure 4: Boxplot for cholera deaths

Additionally, histogram graphs were constructed to examine the normality of data. The distribution of both cholera cases and deaths assumes a normal distribution which shows normality as in the diagrams below.

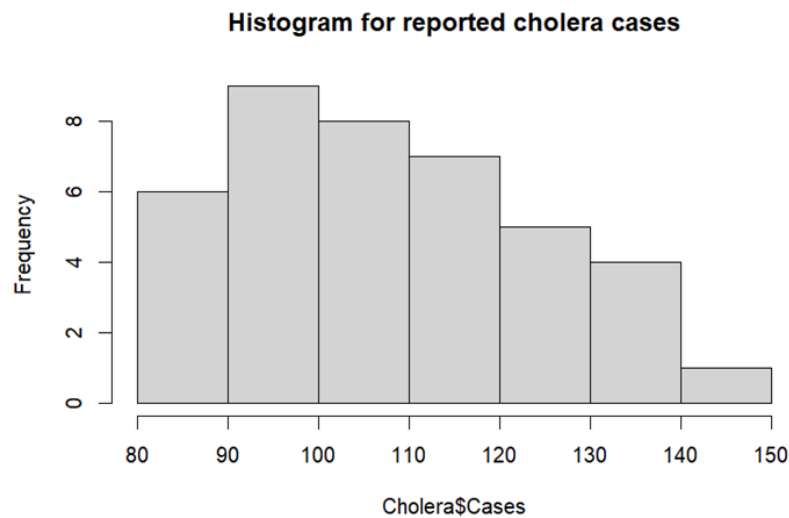


Figure 5: Histogram for cholera cases

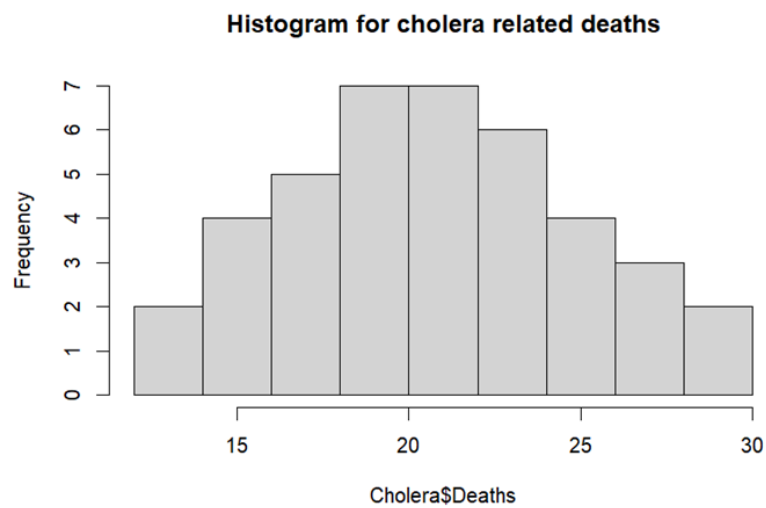


Figure 6: Histogram for cholera deaths

Shapiro-Wilk test was also used to test for normality. Shapiro-Wilk test is defined as a statistical test used to determine if the data is normally distributed or not (Dudley, 2015). This test indicates a normal distribution on the data if the p-value is greater than 0.05. In this case, the p-value for cholera cases was 0.4739 and for cholera deaths was 0.9796, therefore the data was normally distributed.

4.2.2 Stationarity test

The researcher also tested for stationarity using Augmented Dickey Fuller Test (ADF) on cholera cases and cholera related deaths. Augmented Dickey Fuller test is a statistical test that is used to analyze if a time series has a root, making it stationary (R-bloggers, 2021). The p-value for cholera cases was 0.8827 which means the data for cholera cases was non stationary. The p-value for deaths was 0.01 and it showed that the data for cholera deaths was stationary. Since the data for the cholera cases was non-stationary, auto ARIMA was introduced to solve this issue.

4.2.3 Autocorrelation test

ACF and PACF plots were used to test for autocorrelation. The plots showed high correlation on both cholera cases and cholera related deaths. The existence of too many spikes outside the blue line shows the presence of higher correlation which violates one of the time series assumptions. If the data is correlated it also means that it is not stationary. Therefore, auto Arima was performed to solve this issue. The PACF and ACF plots are represented below, providing the presence of autocorrelation in the data.

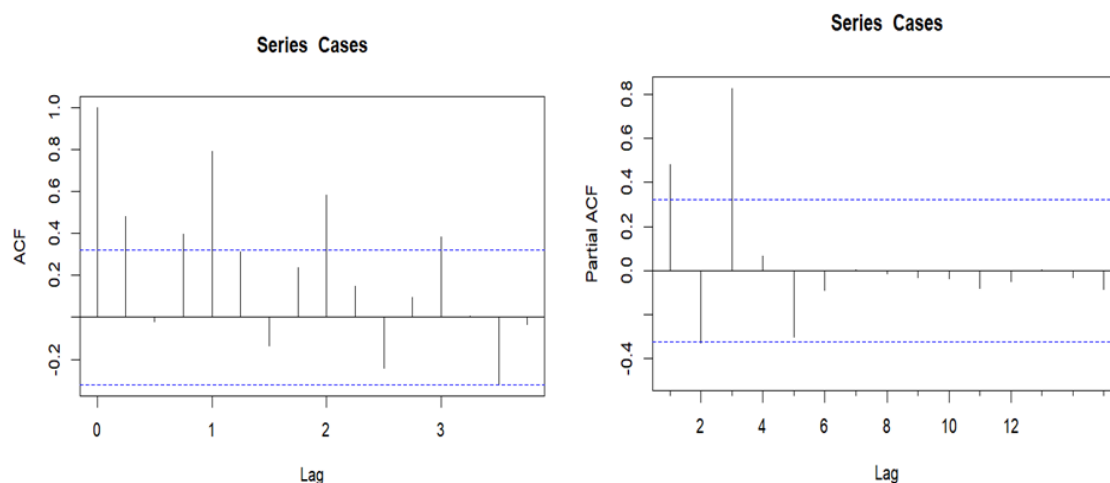


Figure 7: ACF and PACF plots for cholera cases

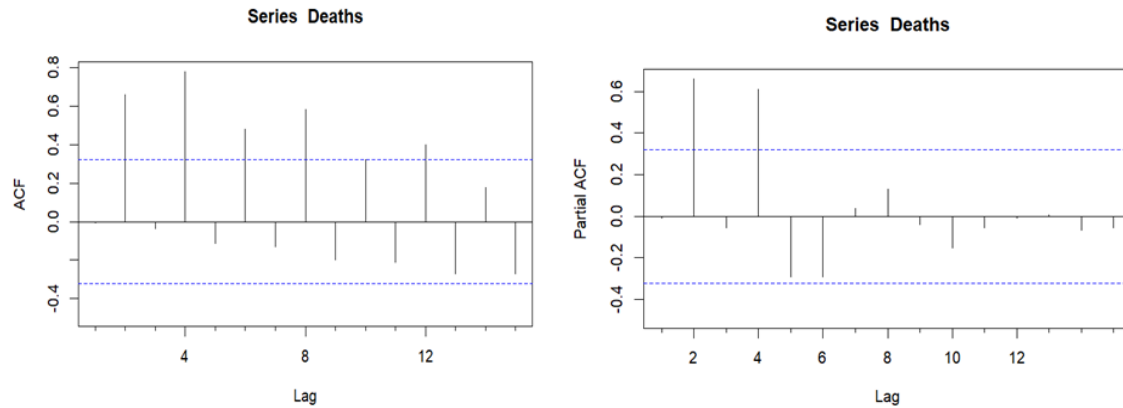


Figure 8: ACF and PACF plots for cholera deaths

4.2.4 Seasonal trends

The data was analyzed to see if there were any seasonal trends on the data. The analysis showed seasonal trends on cholera cases and related deaths. Seasonal trends refer to the patterns or cycles that occur at fixed intervals and they are often related to changes in weather (Mohammed et al. 2019). As shown by the figures below, the cholera cases increased in Q1 and Q4 and also the cholera related deaths are peak in Q2.

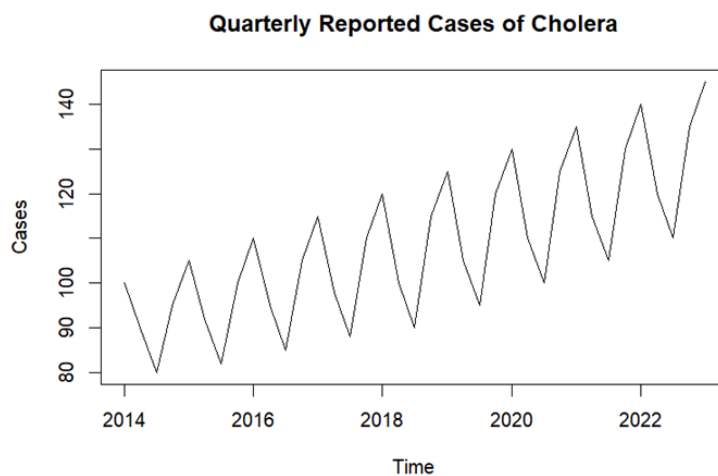


Figure 9: Seasonal trend for cholera cases

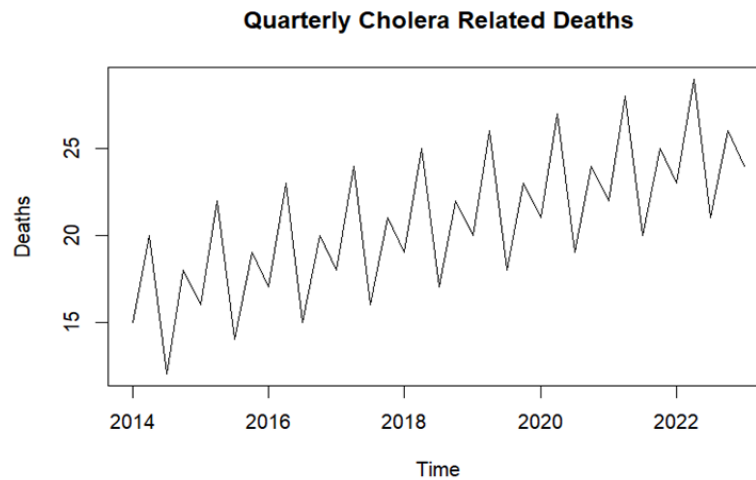


Figure 10: Seasonal trend for cholera deaths

4.3: Model Output

On diagnostic tests, the data showed seasonal trends. The tests also highlighted that the data was autocorrelated as well as non-stationary, therefore, auto-ARIMA was performed to solve this issue in combination with the Akaike Information Criterion (AIC). AIC is a single-number score that is used to select the best model for a given dataset. There were two models that were selected: model 1 for cases and model 2 for deaths. The best model for cases was ARIMA (2,1,0) (0,1,1) [4], and for deaths, it was ARIMA (0,0,1) (0,1,0) [4]. The best models were selected after taking the AIC's lowest scores.

ARIMA (2,1,0) (0,1,1) [4] model shows that (2,1,0) is a non-seasonal component with two autoregressive (AR) terms, indicating that the current value is a function of the previous two values, 1 degree of differencing (I), indicating that the data has a linear trend and no moving average (MA) terms indicated by 0. Moreover, (0,1,1) means that it is a seasonal component with no seasonal AR terms, 1 degree of seasonal differencing (SI), indicating a seasonal trend and 1 seasonal MA term, indicating that the current value is a function of the previous season's error. The seasonal period is 4, indicating quarterly data. Furthermore, ARIMA (0,0,1) (0,1,0) [4] model shows that (0,0,1) is a non-seasonal component with no AR terms, no degrees of differencing (no trend) and 1 MA term, indicating that the current value is a function of the previous error. (0,1,0) is a seasonal component with no seasonal AR terms, 1 degree of seasonal differencing (SI),

indicating a seasonal trend and no seasonal MA terms. [4] is the seasonal period, indicating quarterly data for deaths.

Table 2: Summary for ARIMA (2,1,0) (0,1,1) [4] model

Coefficients	ar1	ar2	sma1
	0.0000	-0.9454	-0.2656
s.e.	0.0714	0.0494	0.1657

AIC=66.59 AICc=68.07 BIC=72.45

Model 1 equation

$$y_t = 0 + 0.0000y_{t-1} - 0.9454y_{t-2} + \varepsilon_t - \varepsilon_{t-4} - 0.2656\varepsilon_{t-5} \dots \dots \dots (4.1)$$

where; $\phi_1 = 0.0000$, $\phi_2 = -0.9454$, $\theta_1 = -0.2656$

Table 3: Summary for ARIMA (0,0,1) (0,1,0) [4] model

Coefficients		ma1	drift
		0.9137	0.2643
s.e.		0.1106	0.0142

AIC=-14.5 AICc=-13.67 BIC=-10.01

Model 2 equation

$$y_t = 0.2643 + \varepsilon_t - \varepsilon_{t-4} + 0.9137\varepsilon_{t-1} \dots \dots \dots (4.2)$$

where; $c = 0.2643$, $\theta_1 = 0.9137$

4.3.1 Residuals

After fitting the ARIMA model, there were residuals. Residuals represent the part of the original data the model was unable to capture. For the cholera cases(model1), the residuals were the differences between the actual cholera cases and the values predicted by the ARIMA (2,1,0) (0,1,1) [4] model. For the deaths(model2), the residuals were the differences between the actual cholera deaths and the values predicted by ARIMA (0,0,1) (0,1,0) [4]. ACF and PACF plots were used to examine if these residuals were correlated. There were too many lags which exceeded the significance bounds and the residuals appeared to be autocorrelated. Upon examining the residuals of the model, autocorrelation was detected. To address this issue, the residuals from each object were converted into time series objects using the ACF and PACF plots. The plots reviewed that the data was uncorrelated. This finding indicates that the residuals are now randomly distributed around the predicted values, supporting the assumption of independence. The uncorrelated residuals are shown by the ACF and PACF plots below.

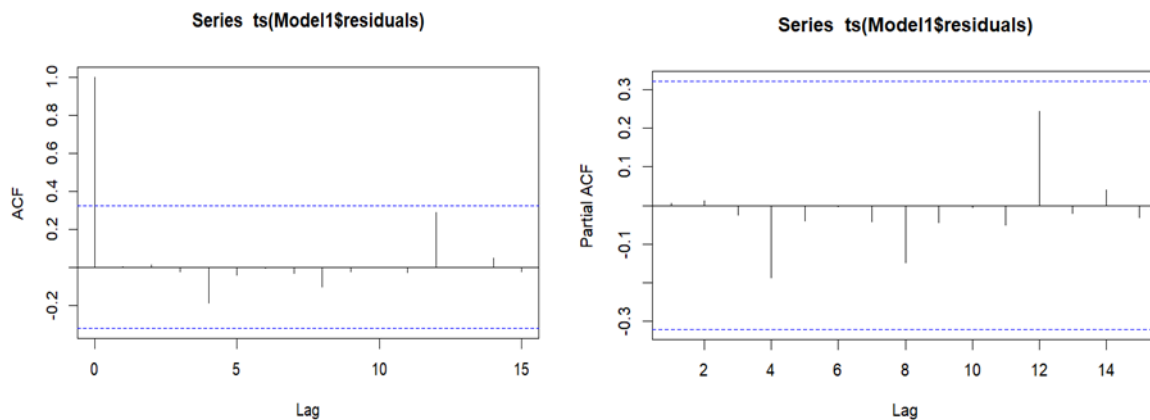


Figure 11: ACF and PACF plots of cholera cases residuals

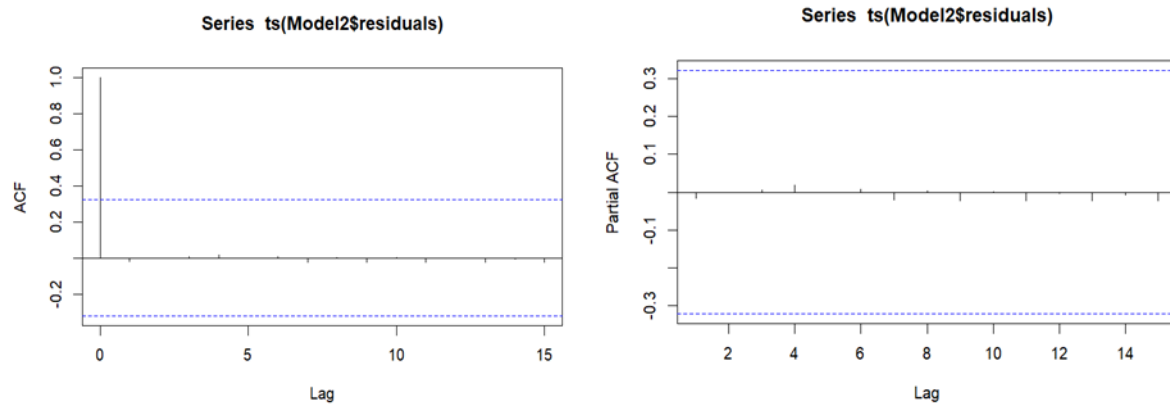


Figure 12: ACF and PACF of cholera deaths residuals

The residuals were plotted overtime to visually analyze their trends, patterns and to identify any other remaining issues on the ARIMA model. The residuals had constant mean and these residuals were closer to zero which indicated that the model was performing well. The plots of residuals over time were presented in the diagrams below, providing a visual representation of the model's effectiveness.

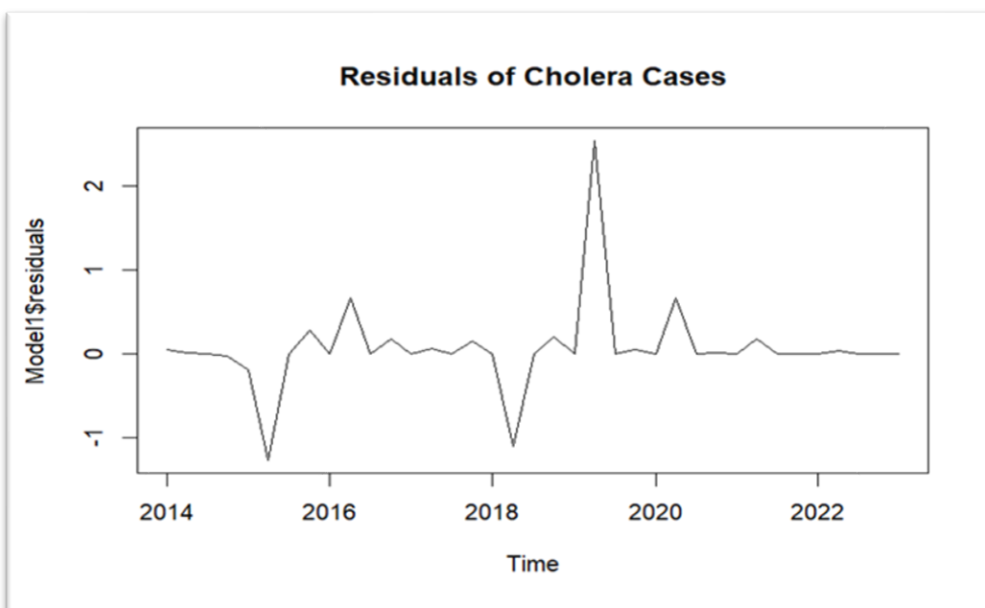


Figure 13: Patterns for cholera cases residuals

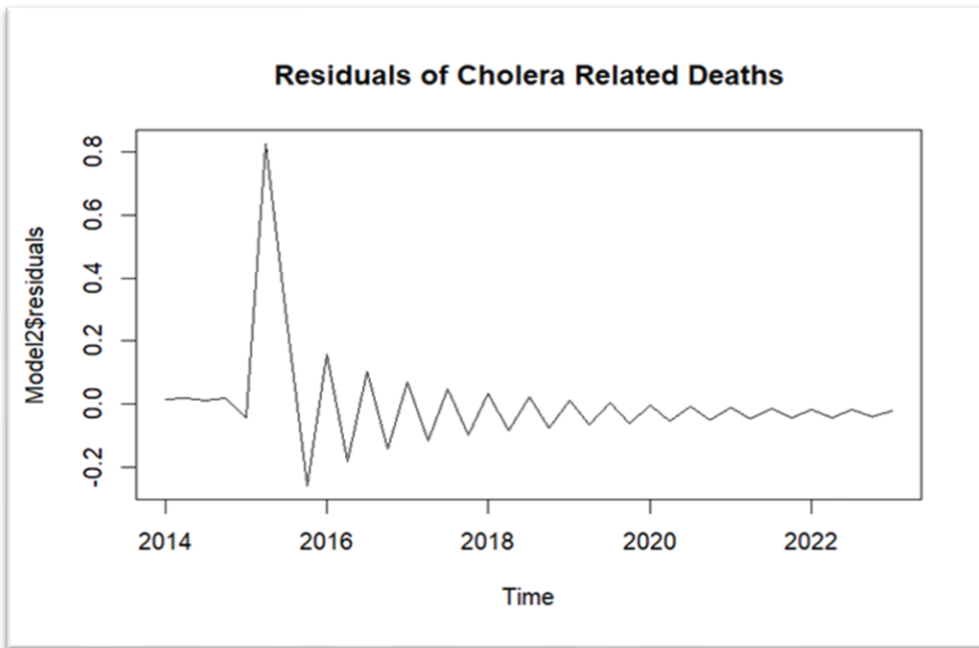


Figure 14: Patterns cholera deaths residuals

4.3.2 Forecasting

Table 4: Predicted cholera cases

	Point	Forecast	Lo 95	Hi 95
2023	Q2	124.9873	123.8106	126.1640
2023	Q3	114.9873	113.3232	116.6514
2023	Q4	139.9990	138.3337	141.6644
2024	Q1	149.9990	148.3325	151.6656
2024	Q2	129.9753	127.3872	132.5633
2024	Q3	119.9753	116.7166	123.2339
2024	Q4	144.9975	141.7344	148.2605
2025	Q1	154.9975	151.7301	158.2649
2025	Q2	134.9638	130.6963	139.2314
2025	Q3	124.9638	119.8895	130.0381
2025	Q4	149.9954	144.9116	155.0791
2026	Q1	159.9954	154.9022	165.0885
2026	Q2	139.9529	133.7864	146.1193
2026	Q3	129.9529	122.8740	137.0317
2026	Q4	154.9928	147.8975	162.0881

Table 5: Predicted cholera deaths

	Point	Forecast	Lo 95	Hi 95
2023	Q2	30.03832	29.68905	30.38758
2023	Q3	22.05702	21.58399	22.53005
2023	Q4	27.05702	26.58399	27.53005
2024	Q1	25.05702	24.58399	25.53005
2024	Q2	31.09533	30.50733	31.68334
2024	Q3	23.11404	22.44507	23.78300
2024	Q4	28.11404	27.44507	28.78300
2025	Q1	26.11404	25.44507	26.78300
2025	Q2	32.15235	31.39770	32.90701
2025	Q3	24.17105	23.35174	24.99037
2025	Q4	29.17105	28.35174	29.99037
2026	Q1	27.17105	26.35174	27.99037
2026	Q2	33.20937	32.31872	34.10002
2026	Q3	25.22807	24.28201	26.17413
2026	Q4	30.22807	29.28201	31.17413

As shown in figure 4 and 5 above, the predicted values reveal an increase in seasonality for both cholera cases and cholera related deaths. Notably, the peak values for cholera cases occur in the first and fourth quarter whereas peak values for cholera deaths occur in the second quarter. The seasonality of the cholera cases and related deaths is also shown by the diagrams below.

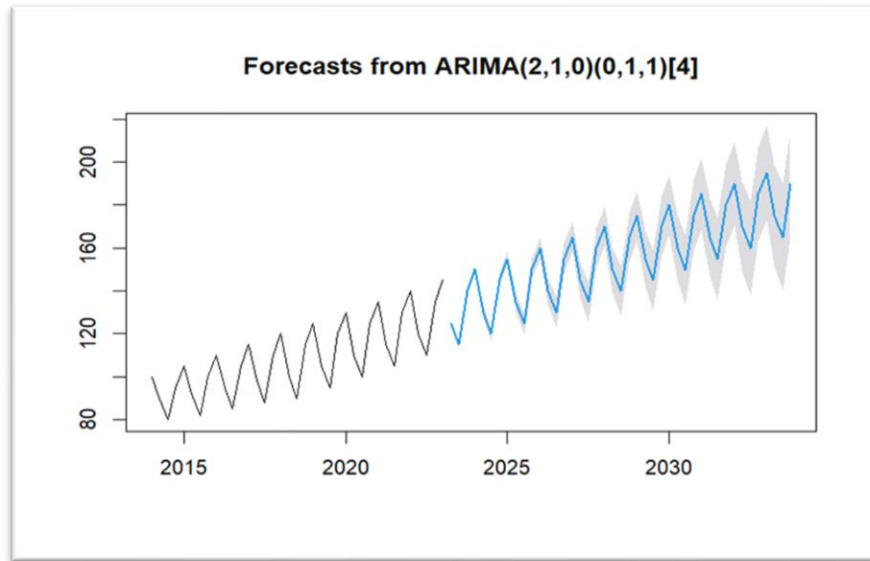


Figure15: Seasonal pattern for predicted cholera cases

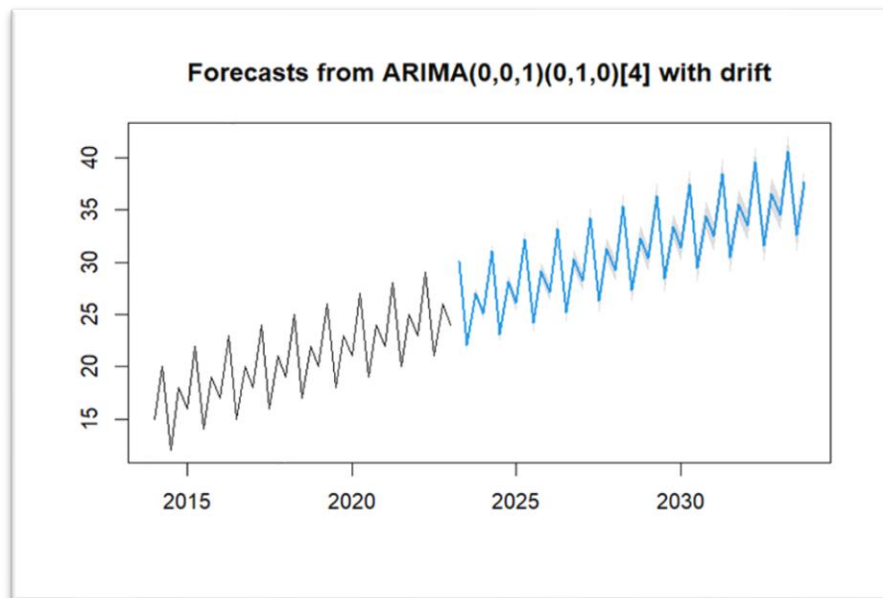


Figure16: Seasonal pattern for predicted cholera deaths

4.4: Model Validation

Parameter estimation for ARIMA (2,1,0) (0,1,1) [4] model

Training set error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.07056364	0.5315115	0.2090516	0.06136774	0.2054066	0.04757726	0.004932808

On model 1, the MAPE was 20.54% which highlighted that the model's forecasts differed from the actual values by about 20.54%. Since MAPE value below 25% is generally considered to represent good predictive accuracy, therefore the calculated MAPE was considered good. The low MAE value of 0.2090516 showed that the magnitude of prediction errors is relatively small. The RMSE value indicated that the size of model's prediction errors was around 0.5315115. The low MAE, MAPE and RMSE values indicated that the model was able to make accurate forecasts with relatively small prediction errors compared to actual values, therefore this model appeared to be good fit for the data.

Parameter estimation for ARIMA (0,0,1) (0,1,0) [4] model

Training set error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.002677069	0.1630843	0.08387301	0.05224099	0.4187484	0.07908027	-0.01712566

For model 2, the MAPE was 41.87%, showing that the model's forecasts differ from the actual values by about 41.87%. This suggests that the model's predictive accuracy was on the higher side and needed some improvements. Despite this, the MAE of 0.08387301 and RMSE of 0.1630843 indicated that the magnitude of prediction errors was relatively small. While the low RMSE and MAE values suggested that the model was capable of making accurate forecasts with small typical prediction errors, the higher MAPE value indicated that there may be errors in some cases.

Furthermore, Ljung-Box was also used for model validation by assessing the residuals for autocorrelation. The results showed that all the p-values exceeded 0.05 for both model1 and model2 ARIMA models, indicating that the residuals are not significantly correlated. The p-values are shown below.

4.5: Conclusion

In conclusion, this chapter has presented the results of the descriptive statistics, diagnostic tests, and model validation for the ARIMA models predicting cholera cases and deaths. The findings indicate that the models are reliable and accurate, with good predictive performance and small prediction errors. The residuals were found to be uncorrelated and randomly distributed, supporting the assumption of independence. The Ljung-Box test confirmed the absence of significant autocorrelation in the residuals. Overall, the results suggest that the ARIMA models are suitable for forecasting cholera cases and deaths, and can be used to inform public health interventions and resource allocation.

CHAPTER 5: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

This final chapter synthesizes the findings of the research, drawing conclusions and impact from the analysis. The chapter summarizes the main discoveries, highlighting the key results and trends that emerged from the data. Based on these findings, conclusions are drawn regarding the predictive model's performance.

5.1 Summary of Findings

The research study was to analyze cholera case data in Bindura District the quarterly data to identify long trends, to investigate the seasonal patterns of the cholera outbreaks and to develop and forecast future trends of cholera cases and deaths. After analyzing the cholera cases and making some predictions with ARIMA (2,1,0) (0,1,1) [4] model which appeared to be a good fit for the data, the continuous seasonal pattern of increased cholera cases was observed in Q1 and Q4. This surge cases are attributed to the high temperatures during these quarters which promotes the growth rate of *Vibrio cholerae*, the bacterium that causes cholera. Furthermore, high temperatures lead to water scarcity, promoting people to use water contaminated sources, thereby increasing the risk of cholera transmission. Additionally, the mobility and gathering of people during these periods may contribute to the spread of the disease, and there is more likely to be high rainfalls in Q1 and Q4 which contaminate water sources, leading to increased cholera cases.

The cholera deaths were forecasted using ARIMA (0,0,1) (0,1,0) [4] model and continuous seasonal pattern of peak values was in Q2. The forecasting results revealed a discordant trend between cholera cases and deaths. While cholera cases are predicted to increase, cholera deaths are not expected to follow a similar pattern, suggesting a decoupling of the cholera cases and deaths

5.2 Conclusions

All the three objectives were fulfilled by the analysis. Clearly, the results indicated that the ARIMA (2,1,0) (0,1,1) [4] model for cholera cases was the best model and captured accurate results, therefore ARIMA model is good for forecasting cholera outbreaks. ARIMA (0,0,1) (0,1,0) [4]

model for deaths suggested that the model was able to make accurate predictions. These findings underscore the importance of implementing targeted public health interventions to mitigate the impact of seasonal cholera outbreaks in Bindura.

5.3 Recommendations

- Prepare for seasonal trends in cholera cases and deaths by increasing public health measures during peak seasons. These measures include oral cholera vaccines, access to safe drinking water and basic sanitation.
- Consider applying ARIMA models to other regions with similar characteristics to enhance public health preparedness and response.
- Conduct regular public awareness campaigns to educate communities about cholera prevention, symptoms and treatment.
- Strengthen disease surveillance systems to detect cholera outbreaks early, enabling prompt response and intervention. Surveillance systems are mechanisms used to monitor, detect and respond to public health events such as environmental infectious disease outbreaks (Centers for Disease Control and Prevention, 2019). These systems are important for disease prevention, control and management (WHO, 2018).

5.4: Areas of Further Study

This study focused on ARIMA model only but more complex mathematical models like machine learning algorithms to predict outbreaks can be developed. ARIMA model can be compared with other models or machine learning algorithms. There is also need to explore the relationship between climate change and cholera outbreaks in the district.

5.5: Chapter Summary

This chapter provided the summary of findings, conclusions and recommendations. The ARIMA models suggested that they were good for forecasting cholera cases in Bindura District and is recommended to use it for future outbreaks on other regions with same characteristics.

REFERENCES

- Ali M, Nelson AR, Lopez AL, Sack D. (2015): Updated global burden of cholera in endemic countries.
- Aschengrau, A., & Seage, G. R. (2013). Essentials of epidemiology in public health (4th ed.). Jones & Bartlett Publishers.
- Ayling, S., Milusheva, S., Maidei Kashangura, F., Hoo, Y. R., Sturrock, H., & Joseph, G. (2023). A stitch in time: The importance of water and sanitation services (WSS) infrastructure maintenance for cholera risk. A geospatial analysis in Harare, Zimbabwe.
- Brockwell, P. J., & Davis, R. A. (2016). Introduction to Time Series and Forecasting. Springer.
- Burnham, K. P., & Anderson, D. R. (2019). Model selection and multimodel inference: A practical information-theoretic approach. Springer.
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 26(3), 1-22. Retrieved from (link unavailable)
- Hyndman, R.J. and Athanasopoulos, G. (2014) Forecasting: Principles and Practice. Springer, New York.
- Hyndman, R. J. (2019). Statistical forecasting: A review of the literature. International Journal of Forecasting, 35(1), 15-27.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer. ISBN 978-1071614174.
- Mohammed A. Al-Hnaity and Mohammed A. Al-Smadi, (2019) "Time Series Analysis and Forecasting" by, London: Routledge, pp. 1
- R-bloggers. (2021, April 12). Augmented Dickey-Fuller (ADF) Test in R
- Richard M. Dudley (2015). "The Shapiro-Wilk and related tests for normality" (PDF). Retrieved 2022-06-16.
- Sack, D.A. and Sack, R.B. (2018) 'Cholera vaccines: WHO position paper', Weekly Epidemiological Record, 93(22), pp. 237-248

Sack, D.A. (2020) Cholera: A biography. Oxford University Press.

Wagner, J. (2019). Stationarity and differencing.

¹ Wei, W. W. S. (2014). Applied Time Series Analysis. New York: Wiley. (pp. 150-170)

World Health Organization (2022) Cholera vaccines: WHO position paper. Available at: (link unavailable) (Accessed: 25 May 2024)

World Health Organization (WHO). (2018). “Drinking water.” http://whqlibdoc.who.int/publications/2008/9789241563673_part3_eng.pdf

World Health Organization. (2020). Cholera - Global Situation. Weekly Epidemiological Record.

World Health Organization. (2019). Cholera - Global Situation. Weekly Epidemiological Record

World Health Organization (WHO). (2023), <https://www.who.int>

APPENDICES

The whole code used in the data analysis and model building.

```
1 Cholera<-read.csv(file.choose(),header = T)
2 library(tseries)
3 library(forecast)
4 library(ggplot2)
5 attach(Cholera)
6 Cases<-ts(Cholera$Cases,start = 2014,end=2023,frequency = 4)
7 print(Cases)
8 Deaths<-ts(Cholera$Deaths,start = 2014,end=2023,frequency = 4)
9 print(Deaths)
10 summary(Cholera)
11 plot(Cases,main='Quarterly Reported Cases of Cholera')
12 plot(Deaths,main="Quarterly Cholera Related Deaths")
13 adf.test(Cases)
14 adf.test(Deaths)
15 Model1=auto.arima(Cases,ic="aic",trace = TRUE)
16 summary(Model1)
17 Model2=auto.arima(Deaths,ic="aic",trace = TRUE)
18 summary(Model2)
19 acf(ts(Model1$residuals))
20 acf(ts(Model2$residuals))
21 pacf(ts(Model1$residuals))
22 pacf(ts(Model2$residuals))
23 plot(Model1$residuals,main="Residuals of Cholera Cases")
24 plot(Model2$residuals,main='Residuals of Cholera Related Deaths')
25 prediction=forecast(Model1,level = 95,h=43)
26 print(prediction)
27 plot(prediction)
28 prediction2=forecast(Model2,level = 95,h=43)
29 print(prediction2)
30 plot(prediction2)
31 Box.test(prediction$residuals,lag = 5,type = "Ljung-Box")
32 Box.test(prediction$residuals,lag = 15,type = "Ljung-Box")
33 Box.test(prediction$residuals,lag = 20,type = "Ljung-Box")
34 Box.test(prediction2$residuals,lag = 5,type = "Ljung-Box")
35 Box.test(prediction2$residuals,lag = 15,type = "Ljung-Box")
36 Box.test(prediction2$residuals,lag = 20,type = "Ljung-Box")
```