BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

Decision Making Modelling for Covid-19 Cases Using Time Series Models the Case of Bindura Provincial Hospital.

SUBMITTED BY:

MUPAGA MYLORD (B192749B)

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE BACHELOR OF SCIENCE HONORS DEGREE IN STATISTICS AND FINANCIAL MATHEMATICS (HBScSFM).

SUPERVISOR: DR T W MAPUWEI

JUNE 2023

DECLARATION OF AUTHORSHIP

I declare that this research project herein is my own original work and has not be copied or extracted from previous sources without due acknowledgement of the sources.

MUPAGA MYLORD 09/06/2023

Name of student **Signature** **Date**

APPROVAL FORM

The undersigned certify that they have read and recommended to Bindura University of Science Education for acceptance of a project entitled **"Decision making modelling for covid-19 cases using time series models the case of Bindura Provincial Hospital".**

Submitted by **MUPAGA MYLORD** in partial fulfilment of the requirements for the Bachelor of Science Honors degree in Statistics and Financial Mathematics.

MUPAGA MYLORD		09/06/2023
Name of student	**Signature**	**Date**
		**12/06/2023**
DR T W MAPUWEI	…….....	…….....
Name of Supervisor	**Signature**	**Date**
Dr. M. MAGODORA	……	……
Name of Chairman	**Signature**	**Date**

RELEASE FORM

Registration Number: B192749B

Dissertation Tittle: **Decision Making Modelling for Covid-19 Cases Using Time Series Models the Case of Bindura Provincial Hospital.**

Year Granted: **2022-2023**

Authority is given to the Bindura University of Science Education Library and the department of Statistics and Mathematics to produce copies of this Dissertation of academic use only.

Signature of student..... 

Date signed… 09/06/2023

DEDICATION

This dissertation is dedicated to my lovely brother Macdonald Mupaga for his sacrifice towards my education. I deprived him of his luxurious life because of the strained budgets after enrolling for this degree program.

ACKNOWLEDGEMENTS

I wish to express my deepest appreciation to my supervisor, Dr T W Mapuwei for supervising me and making valuable contributions in the realization of the success of the project. My special thanks goes to the Ministry of Health (Bindura district hospital) for providing me data. I would also want to express my heartfelt thanks to my friends who have always supported and advised me on academic issues. My lecturers cannot be left out for the assistance they have been giving me during this research. Last but not least, I would like to express my gratitude to my family for the support they gave me during this research, may God bless you abundantly by allowing you to achieve all your aspirations and ambitions.

ABSTRACT

Modelling of communicable diseases is vital to the Public Health Department. COVID 19 is an infectious disease that needs forecasting in order to see the pattern of the epidemic at particular periods. Frequent outbreaks are a serious problem being faced by Health Authorities. It is therefore the aim of the study to come up with a statistical model for COVID-19 situation in Bindura. Data from Bindura district hospital was used to fit the SES model. It looked at exploratory data analysis of the COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023), to fit appropriate forecasting model to the daily cases in Bindura COVID19 data and to predict the future COVID 19 cases using the identified SES model. There is a sharp increase in COVID-19 cases around November, December in 2021 and 2022. The Ministry of Health may advice the general public to adhere to some preventive measures so as to reduce the rate of spread of the disease. These preventive measures may be enforced through a strictly monitored lockdown and discourage large crowd gatherings. Virtual meetings and lessons should replace face to face meetings and lessons so as to reduce direct contact from person to person.

Contents

DECLARATION OF AUTHORSHIP ii

APPROVAL FORM iii
RELEASE FORM iv
DEDICATION v
ACKNOWLEDGEMENTS vi
ABSTRACT vii
LIST OF ABBREVIATIONS xi
CHAPTER 1 1
INTRODUCTION AND BACKGROUND OF THE STUDY 1
1.0 Introduction 1
1.1 Background 1
1.2 Problem Statement 2
1.4 Research Questions 3
1.4.1 What is the COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023)? 3
1.4.2 What is an appropriate forecasting model to the daily cases in Bindura COVID19 data? 3
1.5 Assumptions 4
1.6 Justification of the Research 4
1.7 Significance of the Research 5
1.8 Delimitations of Research 5
1.9 Limitations of the Research 5
1.9.1 Financial Resources 6
1.9.2 Time 6
1.10 Definition of Terms 6
1.11 Dissertation Outline 7
1. 12 Chapter Summary 7
CHAPTER 2 9
LITERATURE REVIEW 9
2.0. Introduction 9
2.1 Time Series Methods and Data 9
2.2. Time Series Plots 10
2.3. Time Series Method and Model Development 11
2.4. Stationarity 13
2.5. Implications of COVID 19 restrictions on the spread of the pandemic. 13

2.6 Literature on Nonlinear Models 13
2.7. Proposed Method and Knowledge Gap 14
2.8 Chapter Summary 14
CHAPTER 3 15
METHODOLOGY 15
3.0. Introduction 15
3.1 Research Design 15
3.2 Study Setting 16
3.3 Population and Sampling Techniques 16
3.3.1 Population 16
3.3.2 Sample Sizes 16
3.3.3 Sampling Methods 16
3.4 Inclusion and Exclusion Criteria 17
3.4.1 Inclusion 17
3.4.2 Exclusion 17
3.5 Methodology and Data Collection Methods 18
3.5.1 Case Study 18
3.5.2 Experimental Research 19
3.6 Methods of Data Analysis 20
3.6.1 Data Source 20
3.6.2 Time Series Plot 20
3.6.3 Simple Exponential Smoothing (SES) 20
3.6.4 Data Analysis 21
3.7 Ethical Considerations 21
3.8 Validity and Reliability 21
3.9 Chapter Summary 22
CHAPTER 4 23
DATA ANALYSIS, FINDINGS AND DISCUSSION 23
4.0 Introduction 23
4.1 Exploratory Data Analysis 23
4.2 The COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023). 23
4.3 To fit an appropriate forecasting model to the daily cases in Bindura COVID19 data. 25

4.4 SES MODEL USING ALPHA 0.2 26
Figure 4.4: model summary when alpha=0.2 26
Figure 4.5: Forecast from Simple Exponential Smoothing (SES) 26
Figure 4.6: model summary when alpha=0.05 27
Figure 4.7: Forecast from Simple Exponential Smoothing 27
Table 4.5 Models results. 28
4.7 Model Evaluation 28
4.7.1 To predict the future COVID 19 cases using the identified forecasting model 29
Figure 4.8: Shows the predicted number of cases using test data set 29
4.5 Chapter Summary 30
CHAPTER 5 31
SUMMARY, CONCLUSIONS AND RECOMMENDATIONS 31
5.0 Introduction 31
5.1 Conclusions 31
5.1.1 The COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023) 31
5.1.2 To fit an appropriate forecasting model to the daily cases in Bindura COVID19 data 31
5.1.3 To predict the future COVID 19 cases using the identified forecasting model 32
5.2 Recommendations 32
5.3 Recommendations for Future Research 33
References 34

LIST OF ABBREVIATIONS

ACF	Autocorrelation function
AIC	Akaike Information Criteria
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving average
MA	Moving Average
MAPE	Mean Absolute Percentage Error
MSE	Mean Square Error
PACF	Partial Autocorrelation Function
RMSE	Root Mean Square Error
DES:	Double exponential smoothing
OLS:	Ordinary least Square
SES:	Single exponential smoothing
WHO:	World Health Organization

CHAPTER 1

INTRODUCTION AND BACKGROUND OF THE STUDY

1.0 **Introduction**

Firstly, coronavirus disease of 2019 (COVID-19) caused by infection with the SARS-CoV-2 virus was first identified in Wuhan City, Hubei Province, China in December 2019 WHO (2020)). The disease has now spread to 216 countries and territories around the world with over 22.5 million confirmed cases and over 790,000 deaths globally as at 19 August 2020 WHO (2020). COVID-19 symptoms, which usually appear 2-14 days after exposure to the virus, include fever or chills, dry cough, tiredness, shortness of breath and sometimes dyspnea WHO (2020). The reverse transcription polymerase chain reaction (RT-PCR) is the gold standard for laboratory diagnosis of SARS-CoV-2 infection MoHCC (2020). Zimbabwe has not been spared by COVID-19. The first COVID-19 case in Zimbabwe was reported on 21 March in the resort town of Victoria Falls MoHCC (2020. By 31 March, seven more people had tested positive, with 1 reported death MoHCC (2020). There was a steady increase in number of cases in the months April to June. In July a surge in cases was reported from 3,659 on 1 August 2020 to 5,378 on 18 August 2020 MoHCC (2020). On 18 August 2020, there are 141 reported COVID-19 related deaths MoHCC (2020).

1.1 Background

In the late December 2019, a new (novel) was identified in China causing severe disease including pneumonia. It was named corona-virus. It spread at an alarming rate in various countries at a fluctuating behavior that resemble a time series pattern. The fluctuations in the spread of the virus have caused problems in the day-to-day operations since the disease spreads at a faster rate in crowded places. People have to go to work places, churches, rallies schools and even watch soccer in stadiums where they would be overcrowded. This on its own caused COVID-19 to be a state of emergency since people had to adopt the new normal way of performing their duties through avoiding gatherings like funerals, churches, rallies, etc. On March 11th, 2020, World Health Organization (WHO) declared the 2019 novel coronavirus as a global pandemic.

In response, many countries have implemented measures such as self-isolation and social distancing in order to prevent further spread, consequently flattening the epidemic curve, which could prove crucial in maintaining health services to patients most in need of care for COVID-19. The ability to identify the rate at which the disease is spreading is crucial in the fight against the pandemic. Being aware of the level of spread at any given point in time has the potential to help governments in public health planning and policy-shaping in order to address the consequences of the pandemic. This study is mainly going to focus on forecasting and predicting behaviour patterns of the spread of the disease in Zimbabwe.

Regardless of health education campaigns, there is still high level of ignorance in some parts of remote communities of Zimbabwe. This is due to some religious believes, fear of the disease as well as denying that the disease is existent among some communities. The information gathered is going to assist epidemiologists with future forecasts on the number of cases and help health authorities enforce relevant preventive measures to reduce the spread of the virus. The potential health threat posed by COVID-19 is very high to the nation as a whole and globally. The study will focus on the behaviour patterns of the virus in Zimbabwe so that the gathered information can be used in risk assessment and management of people with potential exposures to COVID-19.

1.2 Problem Statement

In June 5, 2020, more than 6,603,329 cases have been reported and resulting in more than 391,732 deaths across the world. The first case of COVID-19 pandemic was confirmed in Zimbabwe on March 20, 2020. Consequently, cases increased at a gradual rate from then until an exponential pattern was witnessed resulting in the highest number of cases being recorded per day being 3110 on July 17, 2021. In March, 2020, the government of Zimbabwe announced that schools, sporting events, and public gathering were suspended for 21 days following the WHO recommendations of international lockdowns to curb the spread of the pandemic. Due to the continuation of the outbreak, in April 2020, the president declared a state of emergency in response to the growing number of coronavirus cases with an indefinite end date. Almost all-pandemic disease exhibits their own patterns, which require to be defined by the level of transmission and coverage. The coronavirus (COVID-19) has a fast transmission nature and grow exponentially across the globe. Subsequently, to model the exponential growing rate of the virus, different researchers conducted their study using a linear based time series model (Autoregressive Moving Average) (ARIMA family) model.

However, such linear based time series models cannot handle a data having an exponential growing pattern and results fail to account the dynamics of transmission of the coronavirus. Therefore, ARIMA family models are unable to fit the data well given an exponential growth of COVID-19 transmission. Thus, we should seek to find a model, which can capture a data that have an exponential growing pattern. Therefore, in this study, we use among the common exponential family models such as an Exponential Growth Model, Simple (Single) Exponential Smoothing (SES), and Double Exponential Smoothing (DES) methods. The main motivation of the study is to identify an appropriate model for coronavirus (COVID-19) which has an exponentially increasing pattern. As a result, finding a model that capture such exponentially increasing pattern of the data is the primary motive of this study. The study seeks to find a proper Statistical model that can be used to model future COVID-19 cases, so that health personnel can be prepared for the appropriate preventive measures to be put in place in order to return normalcy in the communities of Zimbabwe.

1.3 Objectives

1.3.1 To examine the COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023).

1.3.2 To fit an appropriate forecasting model to the daily cases in Bindura COVID19 data.

1.3.3 To predict the future COVID 19 cases using the identified forecasting model.

1.4 Research Questions

1.4.1 What is the COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023)?

1.4.2 What is an appropriate forecasting model to the daily cases in Bindura COVID19 data?

1.4.3 How to predict the future COVID 19 cases using the identified forecasting model?

1.5 Assumptions

This study is based on the following assumptions:

1.5.1 The selected sample will be a representation of the Bindura Provincial Hospital.

1.5.2 Data was collected and recorded at uniform time intervals.

1.5.3 The resources were available for the study to be carried out.

1.6 Justification of the Research

The COVID-19 pandemic has demonstrated that there exist risks that have not been accounted for. There is an urgent need to assess quantitatively the risks related to an epidemic or pandemic with similar rigour as is common, for example, for derivatives in quantitative finance. Given the way that viruses are known to mutate and move from animals to humans, it was not a question whether a pandemic could occur, it was only a question when this may happen. Also, in future, new epidemics could emerge at any time and require appropriate long-term risk management. To provide a basis for accurate quantitative risk management for an epidemic or pandemic one needs an accurate understanding of its dynamics. This paper aims to provide such an understanding, suitable for risk management in areas such as health, economics, finance and insurance. It proposes a general time series model for the COVID-19 and similar epidemics. We are living through a crisis that has not been seen in 100 years. This not a crisis for a country and territory, it is a world crisis and is created by the COVID-19 pandemic. The impact and enormous scale of this crisis causing a lot of fear, uncertainty, and anxiety across the whole world. Researchers are trying to understand the different facts with COVID-19 and trying to bring knowledge to us, which may create a better opportunity to face this pandemic. Thus, it led to the researcher to carry out decision making modelling for covid-19 cases using time series models the case of Bindura Provincial Hospital.

1.7 Significance of the Research

The time series models have the ability to capture the decision-making modelling stochastic of an epidemic in all stages of its evolution, in particular, when the number of newly infected is not too large and the number of newly infected fluctuates considerably. The time series models that combine patient and disease characteristics to estimate the risk of a poor outcome from COVID-19 can provide helpful assistance in clinical decision making. Even though there are many advanced data-driven time series methods used to predict the future number of COVID-19 patients, a new and more accurate prediction model is important in the pandemic crisis. The associated contributing factors should be considered to improve model performance. Therefore, the combination of ARM and ARIMA models by selecting the most associated prognostic rules and integrating with ARIMA models could increase the accuracy of decision making in the new cases to better understand the current situation and the progression of COVID-19, which can be easily used by society, organizations, or governments to assess and manage the crisis during the future outbreak. These models are expected to allow for better preparation, organizing hospital resources of further such units and more optimal use of medical personnel and equipment to enhance healthcare decision-making to manage COVID-19 patients in this crisis situation.

1.8 Delimitations of Research

This research will confine itself to a study of Bindura Provincial Hospitals. However, adequate hospitals will be sampled for the purpose of the study to make results more generalizable and it is hoped that the research findings will be utilized to stimulate further research in the other state hospitals to establish whether similar results would be obtained.

1.9 Limitations of the Research

The researcher foretells the following constraints which affect reliability and validity of these findings:

1.9.1 Financial Resources

The financial constraints as the limiting factor as the researcher failed to collect the data from the entire country. The data used has been collected from Bindura provincial hospital meaning this was a hospital based study and not the whole population of the country. This means there could be missing information from the data provided as some people fail to report to the clinics and the hospitals.

1.9.2 Time

The researcher will have a limited time to undertake the research therefore he was not able to cover all hospitals in Zimbabwe Hospitals. The lack of time resulted in the researcher focusing on the hospital. Thus, the little time frame to prepare for the project, Also the time constraints in data collection as researcher will be preparing for examinations, assignments and work. By taking off days from work to concentrate on the research might be of great help.

1.10 Definition of Terms

1.10.1 Decision making (also spelled decision making and decision making) is regarded as the cognitive process resulting in the selection of a belief or a course of action among several possible alternative options. It could be either rational or irrational. The decision-making process is a reasoning process based on assumptions of values, preferences and beliefs of the decision-make Klein, Gary (2018)

1.10.2 Modelling refers to make or construct a descriptive or representational model in this case the decision-making modelling for covid-19 cases using time series models.

1.10.3 Covid-19 (Coronavirus disease 2019) is a contagious disease caused by a virus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first known case was identified in Wuhan, China, in December 2019. The disease quickly spread worldwide, resulting in the COVID-19 pandemic Page J, Hinshaw D, and McKay B (2021).

1.10.4 Time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time Liao, T.

Warren (2015).The are fifteen of time series models and these are Additive white Gaussian noise, Autoregressive fractionally integrated moving average, Autoregressive integrated moving average, Box–Jenkins method, Distributed lag, Error correction model, Gompertz function, Mixed-data sampling, Moving-average model, Nonlinear autoregressive exogenous model, SETAR (model), STAR model, State-space representation, Vector auto regression and Whittle likelihood.

1.11 Dissertation Outline

This study will be organized into five chapters. The specific information contained in these five chapters is listed below.

Chapter One discusses the research background, research questions and objectives, research questions, justification of the study, limitations of the study, significance of the study, delimitations of the study and definition of terms.

Chapter Two provides a review of the literature on decision making modelling for covid-19 cases using time series models the case of Bindura Provincial Hospital.

Chapter Three presents the methodology of the study. It explains the steps involved in developing decision making modelling for covid-19 cases using time series models, sampling and data collection procedures.

Chapter Four presents the results of the statistical analysis.

Chapter Five includes the findings of the study in relation to the hypotheses, and provides managerial implications. The limitations of the study and suggestions for future research are also discussed.

1. 12 Chapter Summary

This chapter outlined on the background and purpose of the search. The study sought to assess, investigate and come up come with solutions of the issues and concerns of emotional labour on employee work outcomes in state hospitals: Case of Bindura Provisional Hospitals. This chapter include background to the study, statement of the problem, objectives of the study, research questions, assumptions, justification of the research, significant of the research, limitations and

delimitations of the study, definition of terms and chapter summary. The next chapter will review of the literature.

CHAPTER 2

LITERATURE REVIEW

2.0. Introduction

In this chapter we will look at some of the articles and literature related to the modelling and forecasting using time series models like ARIMA.SES, Box Jenkins just to mention a few. Some of the previously methods and approaches used in modelling the disease outbreaks in Bindura Zimbabwe will be discussed in detail. ARIMA, SARIMA and ARMA are some of the methods used to analyze data in various literature. The purpose of this literature review is to provide a critical account and critique of the literature relating to the topic of inquiry. In addition, this literature review is done in order to demonstrate why a new research study is required (Helen, 2007).

2.1 Time Series Methods and Data

It is a well-established that mathematical models aid in investigating the transmission dynamics for spread, control and mitigation of infectious diseases (Iboi et al, 2020). In the history of infectious disease modelling, basic reproduction number,R_o, has been thought of as a key variable in measuring the potential for disease spread in a population (P. van de Driessche and J. Watmough, 2014). The basic reproduction number,R_0, is defined as the expected number of secondary infections produced by an index case in a completely susceptible population (van den Driessche and Watmough, 2014). Analysis of the basic reproduction number provides much information about the behavior of an infectious disease over time. In general, a basic reproduction number less than unit ($R_0 < 1$) shows that there is a chance of decline in secondary infections and if R_0 reduces to zero ($R_0 = 0$)then there is a possible halt of a disease (Victor, 2020). If, by contrast, a basic reproduction number greater than unit (R_0>1), then the infected individuals will increase thereby leading to an epidemic state.

The dissertation seeks to come up with a deterministic time series model that can be used to forecast COVID 19 cases in Zimbabwe. ARIMA models function as data-driven and evidence-based

methods to forecast the trends of infectious diseases and formulate public health policies. Some writers used the ARIMA model which can then be used for forecasting daily update of COVID 19 situation on the number of cases, number of recoveries as well as the number of deaths and produced useful forecasts.

2.2. Time Series Plots

There are many publications in the literature on the use of time-series models to predict pandemics. The ARIMA model is widely used for the short-term predictions of infectious disease dynamics and the SARIMA model is used when temporal trends of seasonality exist in the data. In some researches time series analysis was used to construct ARIMA and SARIMA models on the basis of monthly influenza incidence from 2004 to 2011 in four provinces in mainland China. The goal was to predict influenza incidences in 2012. Several research efforts have proposed different time-series models to estimate the spread of COVID 19.

In June 2020, ARIMA was developed to predict the incidences of COVID 19 in India and countries with the highest numbers of confirmed cases, such as in USA, Spain, Italy, France, Germany, China and Iran. Analysis was based on daily COVID 19 data that were collected for the period from 22 January 2020 to 13 April 2020. The ARIMA model was more capable in the prediction of COVID 19 cases compared to other prediction models, including instance support vector machine (SVM) and wavelet neural network (WNN). Existing India COVID 19 data were also used for forecasting new daily confirmed cases using two models, early R and ARIMA. A comparison between the two models showed that the ARIMA model provided better accuracy than that of the early R model.

Liu (2020) states that, "there are four options for the Forecast Approaches parameter, and which will perform best is based on the nature of the time series data in forecasting: Is there an overall non-linear trend like exponential or S-shaped trend? Is there a seasonality added to such trends? Basically, if your data has some overall trend and you want to carry overall trend to the future, not to choose the building model by value because it cannot do extrapolation. Instead, you can start with the two approaches with de-trending. The option building model by value after de-trending considers mostly whether the trend is increasing or decreasing overlay. The option building model by residual after de-trending considers whether the rate of increasing or decreasing is changing recently compared to previous data, and assumes the recent change in the increasing or decreasing is changing recently compared to previous data, and assumes the recent change in the increasing or decreasing

rate will continue to the future, so visually this method can give a more dramatic result than the previous option. When checking the output features of the two approaches, we can see overall patterns are quite similar."

Sulasikin et al (2021), compares the learning models such as Holt's linear trend method, Holt-Winters' additive methods, and ARIMA models. Holt's linear trend method is an extended form of Simple Exponential Smoothing to allow the forecasting of data with a trend. It has a level and trend, but it does not have seasonality. Meanwhile, the Holt-Winters' additive method is an extension of Holt's exponential smoothing, a time series forecasting method for univariate data. Holt (1957) and Winter (1960) extended the approach to capture seasonality. Such a model adds the seasonality factor to the trended forecast, being to model data with a systematic trend or seasonal component. Therefore, selecting appropriate values for p, d and q can be difficult. However, we use the auto ARIMA function from the pmdarima library in Python to do it automatically.

The HoltWinters additive model is an extension of Holts exponential smoothing, a time series forecasting method for univariate data. It is extended so that it adds the seasonality factor to the trended forecast, being to model data with a systematic trend or seasonal component. It is a simple yet powerful and widely used forecasting method which, as mentioned, can cope with trend and seasonal variation and may be used as an alternative to popular Box Jenkins ARIMA family of methods. However, empirical studies have tended to show that the method is not as accurate on average as the more complicated Box Jekins procedure. Exponential smoothing is the procedure of continuously revising a prediction after taking into account the more recent observations matter more than older ones. The HoltWinters additive model is best for data with trend and seasonality that do not increase over time and results in a curved forecast that shows the seasonal changes in the data. Practical issues in implementing the method include the choice of initial values, their sensitivity to unusual events or outliers, the choice of smoothing parameters and the normalisation of seasonal indices.

2.3. Time Series Method and Model Development

Bin Xu et al (2020) published that, to better understand the epidemiological trends and patterns of infectious diseases, it is essential to monitor and analyse the incidence and death on a long-term perspective. Time series analysis builds upon historical data to both extract the underlying time dependent structure and forecast future developments. Autoregressive Integrated Moving Average

(ARIMA) models overcome the limitation of regression analysis and investigate seasonal fluctuation of linear trend and random error. ARIMA has been successfully applied to the analysis of various infectious diseases.

Bin Xu et al (2020) states that a focused intervention strategy targeting specific regions and age groups is essential for the prevention and control of infectious diseases. ARIMA models function as data-driven and evidence-based methods to forecast the trends of infectious diseases and formulate public health policies. The ARIMA model was used for forecasting daily update of COVID 19 situation on the number of cases, number of recoveries as well as number of deaths and produced useful forecasts.

The time-series data is subjected to various processing steps to discover the patterns for better decision making. Apart from pattern discovery and clustering, other important task of time-series data mining include classification, rule mining and summarisation (Fu, 2011). Distance-based clustering, fuzzy c-means (FCM) algorithm, Autoregressive integrated moving average (ARIMA) models and Hidden Markov model (HMM) are few methods adopted for time series clustering and pattern discovery. Time series forecasting depends on the task of analysing past observation of a random variable and generates a model that portrays the underlying relationship and its patterns. Each of the forecasting methods follows four important steps namely, problem definition, information gathering, selecting the best model and forecasting (Hyndman and Athanasopoulos (2020)). The time-series analysis and forecasting for COVID 19 disease is an emerging research paradigm that requires deep knowledge and better experimentations for intercepting the trend and evaluating the predictions.

According to Chimedza et al (ZOU Module) (2004), the aim of a time series is to identify any recurring patterns which could be useful in estimating future values. Time series analysis assumes that the actual values of a random variable in a time series are influenced by a variety of environmental forces operating over time. Time series are influenced by a variety of environmental forces operating over time. Time series analysis attempts to isolate and quantify the influence of these different environmental forces operating on time series into a number of different components and this is achieved through a process known as decomposition of a time series.

Once identified and quantified, these components are used to estimate future values of the time series. We use the assumption in time series analysis to forecast and continue past patterns into the future which is quite applicable in the spread of communicable diseases like COVID 19. Time series models have been used, there is proof that ARIMA model have been trusted more than any model

since it brought more accurate forecasted results than any other model, hence the researcher will use the ARIMA model in this research.

2.4. Stationarity

Seyed, (2011) and Mahsin, (2012), states that all time-series models make use of stationarity data and several researches considered this aspect during their studies and used the Augmented Dicky Fuller (ADF) test to test the presents of a unit root on the COVID 19 data.

2.5. Implications of COVID 19 restrictions on the spread of the pandemic.

Facing COVID 19 without having an effective vaccine many governments panicked and adopted lockdown strategy to prevent the virus from the spread. However, such a strategy hurts the global economy. Sahoo et al (2021) investigated the possibility of containing the virus without lockdown. To this end, mathematical models based on partial differential equations were considered to inspect the effect of proper quarantine with no lockdown on the virus spread. The authors reported that social distancing and proper quarantine of citizens prior to entering their native countries or native states are the best preventive measures in the absence of a vaccine. While Sahoo et al tried to determine general measures to prevent the virus spread; we aim to predict the trend of the virus spread.

2.6 Literature on Nonlinear Models

Almost all-pandemic disease exhibits their own patters, which require to be defined by the level of transmission and coverage. Following the outbreak, the coronavirus (COVID19) have a fast transmission nature and grow exponentially across the globe. Subsequently, to model the exponential growing rate of the virus, different researchers Gautam A, Jha J, Singh AK (2020) conducted their study using a linear based time series model (Auto Regressive Moving Average (ARIMA family) models. However, such linear based time series models cannot handle a data having an exponential growing pattern and results fail to account the dynamics of transmission of the coronavirus. Therefore, ARIMA family models are unable to fit the data well given an exponential growth of COVID-19 transmission. Thus, we should to find a model, which can capture a data that have an exponential growing pattern. Therefore, in this study, we use among the common exponential family models such as an Exponential Growth Model, Simple (Single) Exponential Smoothing (SES), and

Double Exponential Smoothing (DES) methods. The main motivation of the study is to identify an appropriate model for coronavirus (COVID-19) which has an exponentially increasing pattern. As a result, finding a model that capture such exponentially increasing pattern of the data is the primary motive of this study. The main motivation of the study is to identify an appropriate model for coronavirus (COVID-19) which has an exponentially increasing pattern."

2.7. Proposed Method and Knowledge Gap

In the research, the researcher is going to use the SES method since the data examined produces a time series with no trend and seasonality. The data set produces an exponential pattern, and is characterised by some spikes at some points better described as outliers and leverage points. Also, the ARIMA time series analysis model was applied for the decision-making modelling for covid-19 cases several gaps and limitations were identified in this literature review on identifying an appropriate model for (COVID 19).

2.8 Chapter Summary

To sum up, this chapter have just shown that a lot of researches have been done aligned to the emerging of new COVID 19 variants using different models, however there is need to further research on the prediction of COVID 19 cases in communities of Zimbabwe. The researcher seeks to come with a model that can best forecast some cases of COVID 19. The researcher resolved to use time series models as an effective too for modelling data recorded sequentially over time. The objective of this methodology is to capture the temporal dependence between observations through a mathematical model that allows the description of the main characteristics of the data. In general, these records present trends and seasonal components that can be modelled by different statistical techniques.

CHAPTER 3

METHODOLOGY

3.0. Introduction

This chapter will focus on how the study was conducted, the statistical tools, data type, methods, and steps to fit the SES model and computer packages used in this research. Time series forecasting focuses on analyzing past observations of a random variable to develop a model that captures underlying trends and patterns present in the data. The developed model can be used to predict future values of the random variable. This type of analysis is very useful when the underlying data-generation in unknown. There are many models that can be used but this study will mainly focus on one study of using the Simple Exponential Smoothing model and ARIMA Model in Bindura.

3.1 Research Design

Bless et al. (2022) argues that the main aim of the research design is to answer the study objective and the research question. Two research approached, quantitative and qualitative techniques are equally important in statistics. Hopkins (2018) postulates that quantitative research design deals more with numerical analysis and is able to determine the research variable. The author adds that the research designs in quantitative studies are either experimental or descriptive. Thus, the main reason why the researcher chooses quantitative rather than qualitative is based on the nature of the study. The study aims to unpack the decision-making modelling for covid-19 cases using time series models the case of Bindura Provincial Hospital. These subjective experience and narratives cannot be quantified and only qualitative research will provide a nuanced explanation on the phenomenon.

Quantitative research is regarded as the organized inquiry about phenomenon through collection of numerical data and execution of statistical, mathematical or computational techniques. The source of quantitative research is positivism paradigm that advocates for approaches embedded in statistical breakdown that involves other strategies like inferential statistics, testing of hypothesis, mathematical exposition, experimental and quasi-experimental design randomization, blinding, structured protocols, and questionnaires with restricted variety of prearranged answers (Lee, as cited in Slevitch, 2021). From the foregoing, this quantitative study was able to examine the COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023), to fit an

appropriate forecasting model to the daily cases in Bindura COVID19 data and to predict the future COVID 19 cases using the identified forecasting model.

3.2 Study Setting

Bindura Provincial Hospital is located about 88km north of Harare in the outskirts of Bindura town off Mt Darwin road corner Matepatepa & Cleverhill road and near Bindura prison. The study setting refers to the place where the data are collected. In this study, data was collected from Bindura provincial hospital

3.3 Population and Sampling Techniques

3.3.1 Population

The total population being chosen for the study of, Bindura provincial staff of 200 and 3000 COVID19 patients from March 20, 2020 up to March 10, 2023).

3.3.2 Sample Sizes

It refers to the unit from which information is collected and that provides the basis of the analysis. Here it refers to the Bindura provincial hospital health care employees and patients of healthcare sector. The sample of 300 was used.

3.3.3 Sampling Methods

Sampling is a process of selecting a portion of the population to represent the total population and the findings from the sample represent the rest of the group. The advantage of selecting a sample is that it is less costly and time saving than collecting information from a large group of respondents. The selected sample should therefore, have similar characteristics to the population under study to allow generalizability of the results to represent the population (Burns & Grove 2021:365; Polit & Beck 2016:259). In this study both probability and nonprobability sampling were used. The patients

and staff of Bindura Provincial Hospital were sampled using probability sampling and the respondents were selected using non-probability sampling.

Probability sampling technique is a process of selecting respondents into the study that ensures that every member or element of the population has an equal chance of being selected into the study, prevents subjectivity, bias, and allows the results to be generalized to the target population. The probability sampling method does not allow the researcher to intentionally exclude a certain portion of the population. To achieve this probability the sample should be selected randomly (Burns & Grove 2021:297). Non-probability sampling is a process of selecting respondents into the study with less chances of obtaining a representative sample (Burns & Grove 2021:301). Non-probability sampling, by using the convenience sampling technique, was used to select the respondents into the study.

3.4 Inclusion and Exclusion Criteria

Of major concern, Mattson (2022) argues that the criteria should include details of all relevant descriptors necessary for eligibility centers or participants to be included. This criterion includes the set of predefined features utilized to identify participants.

3.4.1 Inclusion

Mattson et al. (2020) argued that inclusion criteria are characteristics that are the participants must have if they are to be incorporated in the study. This principle involves the selection of attributes of subjects for their selection by removing the influence of specific confounding Variables. The inclusion of respondents in this study was based on the permanent employees at Bindura Provincial hospital and patients between 20 March 2020 to 10 March 2023. Being a permanent employee, in this case, means a hospital employee who is not working part-time or anybody doing practical in the hospital.

3.4.2 Exclusion

It involves excluding the characteristics that disqualify prospective subjects from inclusion in the study. The other criteria for this study were to interview only permanent staff, as a matter of fact,

there were some participants who were student nurses and they were willing to be interviewed, but they were excluded since they did not meet the inclusion criteria.

3.5 Methodology and Data Collection Methods

The research utilised case study and experimental research. It used the time series design under experimental research.

3.5.1 Case Study

Case study can be referred to a research approach that is usually used when in-depth inquiry about phenomenon is required in order to discover the causes of underlying principles. Cavaye (2020) referred to case study as case research and as well argued that there is absence of generally acceptable definition of case research, but it is acceptable to give detailed description of case study via the provision of the attributes, advantages and limitations. Crowe et al (2021) gave the following as the process of case study research,

Moreover, Gaille (2018) gave the following as the strengths and weaknesses of case study. It provides confirmable data from direct observations of the specific unit under investigation. Also**,** it can be conducted from a distance meaning that the researcher does not necessarily have to be available in a specific place to conduct case study. Case study is economical when compared to other approaches or strategies because there is insignificant financial implication for reviewing data. On the other hand, case study is structured to reduce the influence of unconscious bias that is believed that everyone will possess through the collection of fact-studded data. However, defining what is fact and what is not is the sole responsibility the researcher gathering data and in essence, the real-time data collection might have personal influence of the researcher by gathering what the researcher wants to see. Also, time consuming this means the data collection process via case study takes larger amount of time when compared with other approaches, because there are massive data collected and the researcher need much time to sort those data. Case study requires little sample size to provide adequate number of data for analysis, because inefficiency might set in if the unit or entity under investigation possessed different demographics or needs that need to be investigated.

3.5.2 Experimental Research

According to Ross and Morrison (2021), the evolution of experimental research can be traced to psychology and education, the emergence of psychology as a novel science in the 1900s structured its research methods on the conventional paradigms that are dependent on experiments to provide principals and laws. Experimental research can be regarded as any investigation performed through a scientific method in which some variables are held constant in order to measure other variables that are under inquiry. Ross et al (2004) added that the following are the major types of experimental research are true experiments, repeated measures, quasi-experimental, time series design and deceptive appearances. The advantages of experimental research are high level of Control this means Experimental research give room for investigator to possess high level of control on variable to gather preferred findings and Specific Results this means as a result of high degree of control of variables the researcher, experimental research provides results that are very specific and consistent.

It is complementary nature this means the experimental research can be paired with other research approaches due its complementary nature. The disadvantages are unrealistic results this means since some variables are subjected to high degree of controls, the data produced might be erroneous and as well look genuine, this is capable of having two effects for the inquirer. Firstly, variables may be influenced in a manner that it will twists the data to a satisfactory and preferred result. It is time consuming this means each variable under experimental research are separated for testing before considering combination of variables and these processes takes longer period of time to achieve. Finally, ethical problems, there exist some variables that cannot be manipulated and a typical example is vaccine made for the control of a particular virus affecting humans and this vaccine are to be test-run on animals. What is to become of those animals if the vaccine performed negatively on those animals? The effectiveness of experimental research undoubtable but possessed ethical or practical problems

3.6 Methods of Data Analysis

3.6.1 Data Source

The secondary data for Bindura Provincial Hospital was collected for the period March 20, 2020 to March 10, 2023 and permission was obtained from the Bindura Provincial Hospital and Bindura University of Science Education authorities.

3.6.2 Time Series Plot

In order to determine the daily COVID 19 new pattern for Bindura Provincial Hospital, time series plots were used. The time series plot for the whole study period using daily number cases was plotted. To develop an appropriate forecast of the data, the data set was divided into in-sample and out-sample forecasts.

3.6.3 Simple Exponential Smoothing (SES)

The SES method was introduced in 1950s and is smoothing time series technique that smoothen data grounded on the exponential window function. The method assigns exponentially decreasing weights over time. According to (Abebe, 2020) the optimal weights are generated according to the data estimations with a given specific weight. The predictions generated using the ES technique weighted averages of past observations. The technique gives decreasing weights to past observations; hence the more recent observation is given higher weight. Reliable estimates are generated by the method (Abebe, 2020). The SES techniques are used when the time series data has no trend and no seasonality, like the Covid 19 data.

For any time period given by *t*, the smoothing function will be:

$$S_t = \alpha Y_t + (1-\alpha)S_{t-1} = S_{t-1} + \alpha(Y_{t-1} - S_{t-1}), t > 0. \quad (3.1)$$

From equation (3.2), Y_t is the current original Covid-19 time series data set time t, S_{t-1} represents the current smoothed time series data obtained after the application of SES onY_t. The smoothing factor is α and its values ranges from 0 to 1, the moving average parameter is estimated by $(1-\alpha)$ while S_{t-1} is the same as the value of Y_t, that is $Y_t = S_t$.

If α is closed to one, it means less smoothing effect and this will allocate greater weight to recent changes in the observations but a smoothing factor that is close to zero means greater smoothing

effect, hence less responsive to recent changes in the time series data. For estimation purposes that is estimating *h*-steps ahead, we can have:

$$Y_{t+h} = S_{t,} h = 1,2,3, \dots, n. \quad (3.2)$$

The general SES forecast function will be given by:

$$Y_{t+h|t} = \sum_{i=1}^{N} (1-\alpha)^{t-i} Y_i + (1-\alpha)^t S_{0,} \ 0 \leq \alpha \leq 1, \quad (3.3)$$

Where $Y_{t+h|t}$ represent *h*-step ahead forecast from original *t* and *h* is the number of periods in the forecast lead-time.

3.6.4 Data Analysis

The data analysis utilised statistical packages R and STATA. The independent variable is age, dependent variables are place of residence, reporting health facility, occupation, date of consultation and date of outcome. The dataset was plotted in graph and determined the rolling statistics, and then the model performed the Augmented Dickey–Fuller test (ADF) test t determine the Test Statistic, p-value, Lags Used, and Number of Observations Used. After that, the model estimated the trends, which was the log of the indexed dataset.

3.7 Ethical Considerations

To ensure the ethical conduct of the study the researcher sought and obtained permission to conduct the study from Bindura University of Science Education and Bindura Provincial Hospital authorities. Privacy and confidentiality are based on the principle of respect. Privacy is the right of an individual to determine the circumstances, time, and extent, type of information to share or withhold from others (Burns & Grove 2021:200; Polit & Beck 2016: 91). In this study, the respondents' privacy was maintained by conducting individualized interviews and omission of personal details in the interview schedule and not being forced to answer questions

3.8 Validity and Reliability

According to Bryan and Bell (2007) reliability is achieved if the research results are repeatable. Reliability is the degree of consistency with which the data-collection instrument produces the same

results every time it is implemented in the same situation or used by different investigators. The data-collection instrument should be accurate and stable to reflect true scores of the attributes under investigation and minimize error (Brink 2016:171; Burns & Grove 2021:396, 399, Polit & Beck 2016:324, 328). To ensure reliability, the researcher pre-tested the interview schedule on Bindura Provincial hospital staff who were not part of the sample. This was done to identify vague, unacceptable questions and consistency of results. Validity is the extent of accuracy of an instrument to measure the construct it is supposed to measure in the context of the concepts/variables being studied (Brink 2016:167; Polit & Beck 2016:329). A structured interview schedule was developed after a review of relevant literature to incorporate and measure important variables in the study. The researcher closely examined the questions in the interview schedule to ensure that they measured the desired variables.

3.9 Chapter Summary

This chapter covered the research design, methodology and data collection methods, population and sampling techniques, inclusion and exclusion criteria, methods of data analysis, validity and reliability and ethical considerations. The next chapter is Data Analysis, Findings and Discussion.

CHAPTER 4

DATA ANALYSIS, FINDINGS AND DISCUSSION.

4.0 Introduction

The previous chapter focused on research methodology which covered research design, population size, sample size and sampling procedure, research instruments, validity, reliability, data collection, presentation and analysis procedures as well as ethical considerations. This present chapter focuses on data presentation analysis and discussion of the research findings.

4.1 Exploratory Data Analysis

The dataset contains covid 19 number of cases recoded in Bindura form different hospitals from the period of March 2020 to march 2023. It contains cases for health workers and cases of patients in general. The time series analysis for the positive cases of covid 19 is plotted against date in which the cases were recorded, that is the day the results were out from the lab. The time series analysis for covid 19 cases for health workers is plotted separately from that of the patients, then a joint plot is plotted to identify the trends from total cases.

4.2 The COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023).

Firstly, Covid 19 affected both health workers and patients. To identify the patterns, the analysis of health workers affected by covid 19 is carried first, then the affected patience and a combination of both cases is carried out. The main aim of this objective is to analyse when the cases when the cases were increasing and decreasing in a course of 3 years and in which seasons.

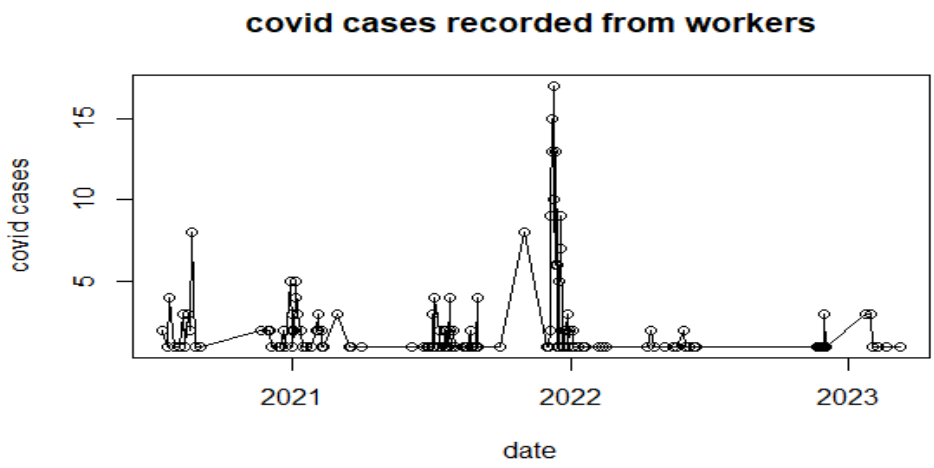

Figure 4.1: Covid 19 Cases Recorded From Workers

The above time series analysis (Figure 4.1) plotted the cases of covid 19 found in health workers. From March 2020 to March 2023. The approximate maximum number of cases report from health workers from 2020 to 2023 is 16 cases, with highest cases reported towards 2022 that is in November and December of 2021. From the Figure, the trends of covid 19 identified suggest that for each year the highest number of cases were recorded towards year end except for 2020 where highest cases were recorded in winter. In 2023 the cases were decreasing, the reason to this could be the vaccination.

Covid Cases Recorded from Patients

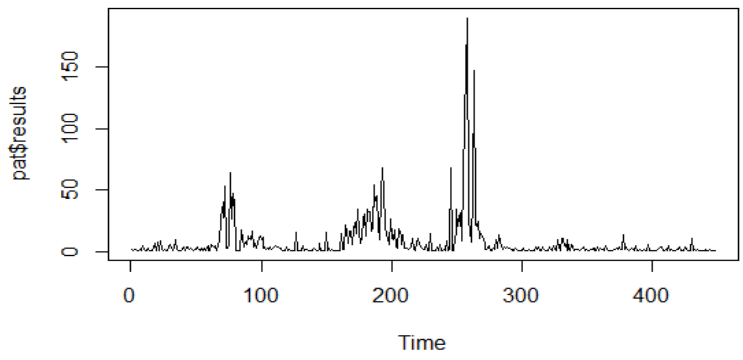

Figure 4.2: Covid 19 Cases Recorded from Patients

Furthermore, the above time series plotted (Figure 4.2) show the trends of covid 19 cases for patients that is the general public. The graph is indicating that there is a long-term increase and decrease in the number of cases. The figure shows daily number of cases with more data points below 50 daily cases. The figure also shows that the highest daily number of cases recorded surpasses 150 cases.

Covid Cases Recoded

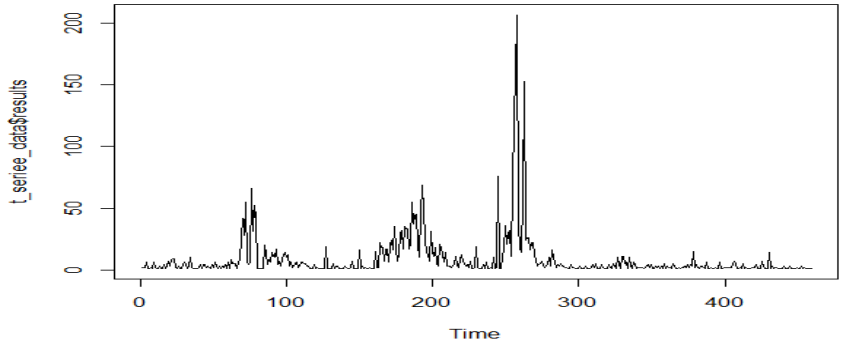

Figure 4.3: Total Covid 19 Cases Recorded

From figure 4.3, the above time series plotted to show the trends of covid 19 total cases for both patients and health workers. From the graph it is clear that the total number of cases recorded for covid 19 forms a trend. The highest total number of cases reported on a single day for both patience and health workers are 200.

4.3 To fit an appropriate forecasting model to the daily cases in Bindura COVID19 data.

To forecast covid 19 cases in Bindura Simple Exponential Smoothing model is used. Before fitting the SES model, the data for total number of cases is split into train and test. The train data is to train the model and test data to evaluate model's performance. The data was split in the ratio of 70:30; where 70 represents the train data and 30 represents the test dataset. The model is first trained as shown in the below figure by setting alpha = 0.2 and the number of steps h = 100.

4.4 SES MODEL USING ALPHA 0.2

```
Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing

Call:
 ses(y = train$results, h = 100, alpha = 0.2)

  Smoothing parameters:
    alpha = 0.2

  Initial states:
    l = 2.09

  sigma:  17.6018

     AIC     AICc      BIC
3745.937 3745.974 3761.504

Error measures:
                     ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
Training set -0.01164654 17.54753 7.63982 -124.0528 153.3566 1.122586 0.398985

Forecasts:
    Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
326       1.332975 -21.22460 23.89055 -33.10587 35.83181
327       1.332975 -21.67133 24.33728 -33.84908 36.51503
328       1.332975 -22.10955 24.77550 -34.51927 37.18522
329       1.332975 -22.53972 25.20567 -35.17717 37.84312
330       1.332975 -22.96228 25.62823 -35.82341 38.48936
331       1.332975 -23.37761 26.04356 -36.45861 39.12456
332       1.332975 -23.78608 26.45203 -37.08331 39.74926
333       1.332975 -24.18801 26.85396 -37.69801 40.36396
334       1.332975 -24.58371 27.24966 -38.30317 40.96912
```

Figure 4.4: model summary when alpha=0.2

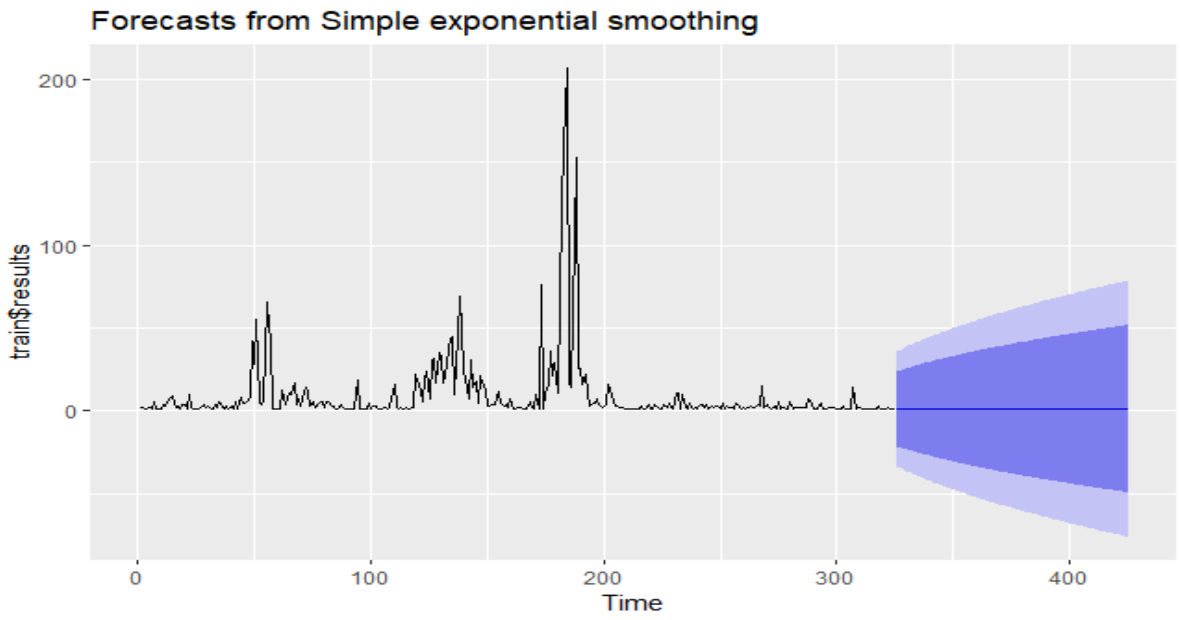

Figure 4.5: Forecast from Simple Exponential Smoothing (SES)

Figure 4.5, the plot shows the results of positive cases on the y axis and x axis time which is representing the steps h. From the time series forecast the purple part represents the future values. The Forecast projects the straight-line estimate into the future, for that verification of seasonal and trend components is to be carried out.

```
Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing

Call:
 ses(y = train.dif, h = 100, alpha = 0.05)

  Smoothing parameters:
    alpha = 0.05

  Initial states:
    l = 0.1

  sigma:  18.7677

     AIC     AICc      BIC
3774.980 3775.018 3782.542

Error measures:
                     ME     RMSE      MAE MPE MAPE      MASE       ACF1
Training set -0.009007037 18.7097 7.065519 NaN  Inf 0.613155 -0.2219381

Forecasts:
    Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
325     -0.04591401 -24.09771 24.00588 -36.82996 36.73813
326     -0.04591401 -24.12775 24.03592 -36.87591 36.78408
327     -0.04591401 -24.15776 24.06593 -36.92180 36.82997
328     -0.04591401 -24.18773 24.09590 -36.96764 36.87581
329     -0.04591401 -24.21767 24.12584 -37.01342 36.92159
330     -0.04591401 -24.24756 24.15574 -37.05914 36.96732
331     -0.04591401 -24.27742 24.18560 -37.10481 37.01298
332     -0.04591401 -24.30725 24.21542 -37.15042 37.05859
333     -0.04591401 -24.33703 24.24521 -37.19598 37.10415
334     -0.04591401 -24.36678 24.27496 -37.24148 37.14965
335     -0.04591401 -24.39650 24.30467 -37.28692 37.19509
336     -0.04591401 -24.42618 24.33435 -37.33231 37.24048
337     -0.04591401 -24.45582 24.36399 -37.37764 37.28581
338     -0.04591401 -24.48542 24.39360 -37.42292 37.33109
```

Figure 4.6: model summary when alpha=0.05

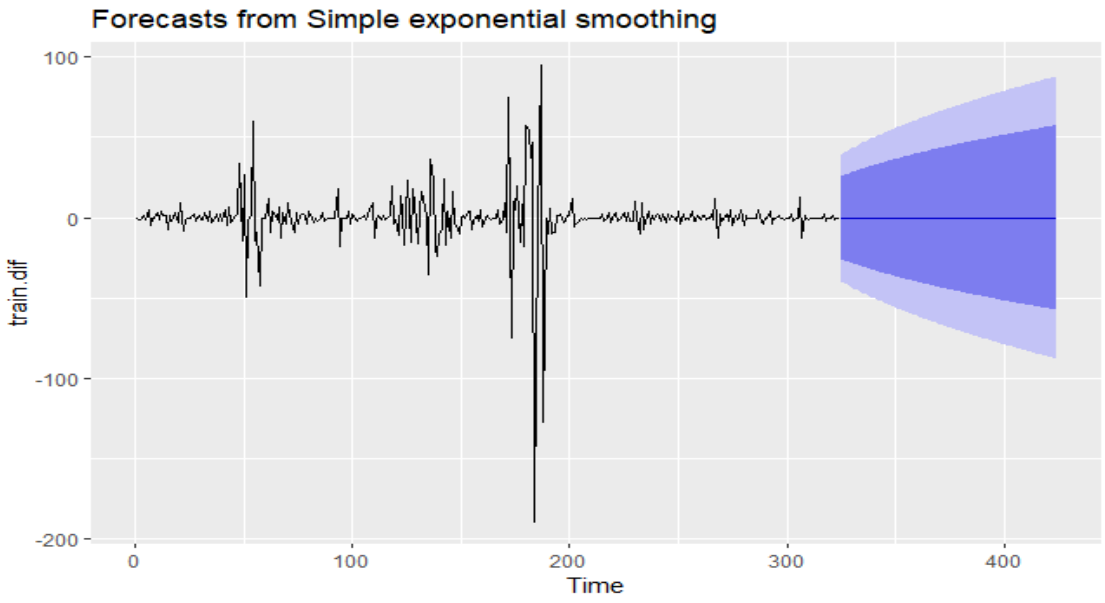

Figure 4.7: Forecast from Simple Exponential Smoothing

The Figure 4.7 forecast plot shows the trained model after changing the level of alpha to 0.05. It is evident from the plot that after changing the value of alpha the model becomes insignificant, that is it over fits the data. From the above figure we conclude that there is no trend and seasonal component.

Comparing the model's summary

Figure 4.4 and 4.6 shows two SES models trained using the partitioned data that is the 70% subset of the Bindura hospital COVID 19 cases data. The aim was to compare, at which level of alpha does the model performs very well, which is the level of alpha at which the model best fits the data and best estimates the covid 19 cases.

Table 4.5 Models results.

	Model 1 (Alpha=0.2)	Model 2 (Alpha=0.05)
AIC	3745.937	3774.980
AICc	3745.974	3775.018
BIC	3753.504	3782.542
Sigma	17.6018	18.7677
ACF1	0.399	-0.2219
MAE	7.64	7.07
RMSE	17.55	18.01

The table above shows model performance at different level of alpha. From the results it is clear that the model trained using alpha= 0.2 is the best. Considering Akaike Information Criterion (AIC), the model with alpha =0.2 is the best model. It is clear from figure 4.4 and 4.6 that model 1 best fits the data and it avoids overfitting the data. The point forecasts, which is the most likely outcome for the next period of COVID 19 cases is positive on model 1 and negative on model 2 signifying that model 2 out performs model 1. Alpha =0.2. The best model uses alpha =0.2 and h =100 with positive point forecasts, therefore model 1 with alpha =0.2 is used to predict future cases.

4.7 Model Evaluation

To evaluate model's performance, test data was used. The reason for evaluation is to see how well our model does performs on the data that it has not seen before.

Table 4.7; model performance

	MAE	RSME	ME	MAPE
Train set	7.64	17.55	-0.0116	153.357
Test set	5.665	9.225	0.0112	126.225

To evaluate our model performance the researcher computed the above metrics for train data and test data to check whether the model is perfect for predicting the future cases of covid 19. The mean absolute error reduced for the test data to 5.665 and root mean squared error to 9.225. The computed metrics provides insights in how well our estimates are. They also show the difference between the actual values and predicted values of the model. To check whether the model is performing well we should reduce the two metrics. We can conclude the model is perfect to use for prediction since we managed to reduce root mean squared error and Mean absolute error for test data.

4.7.1 To predict the future COVID 19 cases using the identified forecasting model.

The model is used to predict the future cases of covid 19. The trained model is used to forecast the number of cases by estimating.

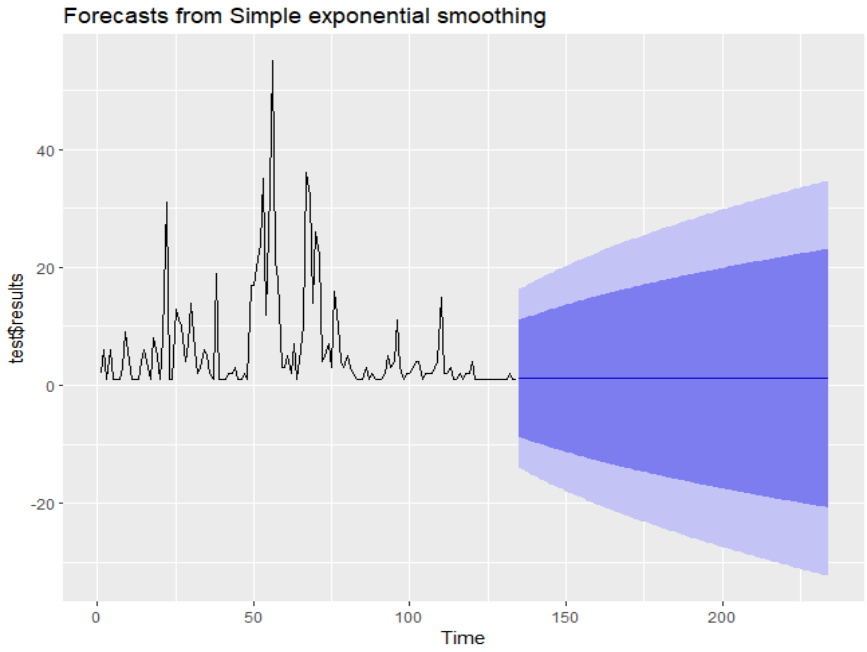

Figure 4.8: Shows the predicted number of cases using test data set

The Figure 4.8 shows the predicted number of cases using the test data set. To check our model performance for predicting the number of covid 19 cases model summary is used.

4.5 Chapter Summary

This chapter looked at data presentation, analysis and discussion. It looked at exploratory data analysis of the COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023), to fit appropriate forecasting model to the daily cases in Bindura COVID19 data and to predict the future COVID 19 cases using the identified forecasting model. The next chapter covers the summary, conclusions and recommendations of the study.

CHAPTER 5

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

The previous chapter looked at data presentation, analysis and discussion of the research findings. It looked at exploratory data analysis of the COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023), to fit appropriate forecasting model to the daily cases in Bindura COVID19 data and to predict the future COVID 19 cases using the identified forecasting model. This present chapter focuses on the summary, conclusions and recommendations of the study.

5.1 Conclusions

5.1.1 The COVID19 patterns for Bindura over the study period (March 20, 2020 up to March 10, 2023)

From March 2020 to March 2023. The approximate maximum number of cases report from health workers from 2020 to 2023 is 16 cases, with highest cases reported towards 2022 that is in November and December of 2021. From the above figure, the trends of covid 19 identified suggest that for each year the highest number of cases were recorded towards year end except for 2020 where highest cases were recorded in winter. In 2023 the cases were decreasing, the reason to this could be the vaccination.

5.1.2 To fit an appropriate forecasting model to the daily cases in Bindura COVID19 data

In regard to Bindura Provincial Hospital's COVID-19 dataset and based on the class of models considered in this study, we can say that for cases with a confirmed diagnosis of COVID-19, the best model corresponds to the Simple exponential smoothing model compared to the ARIMA model due to the nature of the data set where ARIMA needs huge (big data set). To forecast covid 19 cases in Bindura Simple Exponential Smoothing model is used. Before fitting the SES model, the data for total number of cases is split into train and test. The train data is to train the model and test data to evaluate model's performance. To split the data library caTools was used and the data was spit in the ratio 70:30, 70% to train and 30% to test data. Library forecast is used to train the model. The

model is first trained as seen in the below figure by setting alpha = 0.2 and the number of steps h = 100. The plot showed the results of positive cases on the y axis and x axis time which is representing the steps h.

From the time series forecast in chapter 4, the purple part represented the future values. The forecast projected the straight-line estimate into the future, for that verification of seasonal and trend components is to be carried. Forecast plot shows the trained model after differencing the data. It is evident from the plot that after differencing there are negative cases and we concluded that there is no trend and seasonal component.

5.1.3 To predict the future COVID 19 cases using the identified forecasting model

The results the mean error was minimum which suggested that the model best fits our data from plotted graph. The best model used alpha =0.2 and h =100 with positive point forecasts. Though when alpha =0.05 yields better metrics the point forecasts are negative, therefore model 1 with alpha =0.2 is used to predict future cases. The model is used to predict the future cases of covid 19 and the trained model was used to forecast the number of cases by estimating. Shows the predicted number of cases using the test data set. To check our model performance for predicting the number of covid 19 cases model summary was used.

5.2 Recommendations

The future modifications to further improve the predictive accuracy of the models will include the creation of ensembles of the presented models that would combine the best of many worlds in order to reduce the overall error as well as the adoption of multivariate time series modeling that take into account other factors that are either directly or indirectly related to the spread of the pandemic. Another future ambition would be to use some form of transfer learning in order to bring learning's from one country to another in order to know the majority parameters for the actual cause of the spread. In future, researcher can explore some prediction models such as an artificial neural network (ANN), Bayesian networks, and Support Vector Machines (SVM) in COVID-19. This model is also applicable in future pandemics and to predict any type of disease affected patients.

5.3 Recommendations for Future Research

The study was conducted in Bindura provincial hospital set up which makes it hard to generalize its findings. The study therefore recommends that similar studies be executed in other provincial and central hospital nationwide in order to redress this acknowledged limitation.

References

Cavanaugh et al (2021). Reduced Risk of Reinfection with SARSCoV-2 After COVID-19 Vaccination - Kentucky, May-June 2021. MMWR. Morbidity and mortality weekly report, 70(32), 1081-1083.

Centers for Disease Control and Prevention. Coronavirus Disease 2019 (COVID-2019).

De Livera A, Hyndman R, Snyder R. (2021) Forecasting Time Series with Complex Seasonal Patterns Using Exponential Smoothing. Journal of the American Statistical Association; 106(496), 1513–1527.

Domenico B., Marta G., Lazzaro V., Silvia A., and Massimo C. (2020).Application of the ARIMA model on the COVID- 2019 epidemic dataset.Data in brief, 29: 105340.

Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci.(2008). Introduction to Time Series Analysis and Forecasting, 1st publication. A JOHN WILEY . SONS, INC.,PUBLICATION, United States of America.

Edgars P., 2003, Simple and complex market inefficiencies: Integrating efficient markets, behavioral finance and complexity, Journal of Finance, Vol. 12(2), pp 202-224

Fanelli, D., Piazza, F., (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France'. Chaos Solitons Fractals 134, 109761.

G.R. Shinde, A.B. Kalamkar, P.N. Mahalle, N. Dey, J. Chaki, A.E. Hassanien, (2020) Forecasting models for coronavirus disease (covid-19): a survey of the state-of-the-art. SN Comput. Sci. 1(4), 1–15.

Havers et al (2021). COVID-19-associated hospitalizations among vaccinated and unvaccinated adults?

Helen .A (2007). Doing literature review in health and social care. A practical guide, Open University press, McGraw hill, USA. https://www.researchgate.net. Dr Siddaway. What is a systematic literature review and how do I do one?

Hyndman RJ, Koehler AB, Ord JK, Snyder RD. (2022) Forecasting with Exponential Smoothing: The State Space Approach. Berlin Germany: Springer. 372 p.

J.J. LaViola, (2022) Double exponential smoothing: an alternative to Kalman filter-based predictive tracking. Proc. Worksh. Virt. Environ, 199–206.

Keeling and Rohani (2011). Modeling Infectious Diseases in Humans and Animals. Princeton University Press. https://doi.org/10.1515/9781400841035-003

Laith et al (2021). Protection afforded by the BNT162b2 and mRNA-1273 COVID-19 vaccines in fully vaccinated cohorts with and without prior infection. doi:

Lindsey Wang (2021). Increased risk for COVID-19 breakthrough infection in fully vaccinated patients with substance use disorders in the United States between December 2020 and August 2021. Case Western Reserve University, Cleveland, OH, USA

Maleki Mohsen and Mahmoudi Mohammad Reza and Wraith Darren and Pho Kim-Hung. (2020). Time series modelling to forecast the confirmed and recovered cases of COVID-19. Travel Medicine and Infectious Disease; 37: 101742.

Murray J. D. (2001). Mathematical Biology: An Introduction, 3rd edition, volume 17, Springer, New York, U.S.A

Nazim, A., Afthanorhan, A. (2014). A comparison between single exponential smoothing (SES), double exponential smoothing (DES), Holts (Brown) and adaptive response rate exponential smoothing (ARRES) techniques in forecasting Malaysia population. Global Journal of Mathematical Analysis, 2(4), 276-280.

RUEY, S. TSAY.(2005). Analysis of Financial Time Series, Second Edition.John Wiley Sons, Inc., Hoboken, New Jersey.

T.C. COVID et al. , (2020) Characteristics of health care personnel with covid-19-united states, February 12–April 9, 2020. MMWR Morb. Mortal. Wkly. Rep. 2020 69(15), 477–481.

T.T. Le, Z. Andreadakis, A. Kumar, R.G. Roman, S. Tollefsen, M. Saville, S. Mayhew, (2020) The covid-19 vaccine development landscape. Nat. Rev. Drug Discov. 19(5), 305–306.

Usherwood, T. (2021). A model and predictions for COVID-19 considering population behavior and vaccination. Sci Rep 11, 12051 (2021).

van den Driessche P and J. Watmough (2014). Further notes on the Basic Reproduction number, Researchgate.

Vinay Kumar Reddy Chimmula, Lei Zhang.(2020). Time series forecasting ofCOVID-19transmissioninCanadausingLSTMnetworks.Chaos,Solitons and Fractals, 135; 109864

WHO. Coronavirus disease 2019 (COVID-19) situation reports.

World Health Organization (WHO), Coronavirus.(2020).

Yichi L., Bowen W., Ruiyang P., Chen Z., Yonglong Z., Zhuoxun L, Xia J. and BinZ.(2020).MathematicalModelingandEpidemicPredictionofCOVID-19 and Its Significance to Epidemic Prevention and Control Measures.Annals of Infectious Disease and Epidemiology; .

Appendix

```
library(readxl)
workers <- read_excel ("MYLORD MUPAGA/workers.xlsx")

######## exploratory data analysis
###cleaning covid cases for workers
data <- workers[ , colSums(is.na(workers))==0]

#### renaming variables
data$results<-data$`Lab Result for COVID19`
data$date<-data$`Lab Results Date (dd/mm/yyyy)`
########recoding variables
data$results<-ifelse(data$results=="positive", 1,0)

###### drop unnecessary columns from the dataframe
library(dplyr)
library(tidyverse)
library(lubridate)

tm_data <- data %>% select(results, date)
#######
library(data.table)
library(lubridate)

tm <- tm_data %>% group_by(date) %>% summarise(results=n())
##############################################################################
######################
#################### lets load the patients cases data and join the two
patients_cases <- read_excel("MYLORD MUPAGA/patients cases.xlsx")
### cleaning and renaming the data
patients_cases$results<-patients_cases$`Lab results for patients`
```

```
patients_cases$date<-patients_cases$`Results date`
####### dropping na's
patience_data <- drop_na(patients_cases)
####################selecting useful columns
patience_data <- patience_data %>% select(results, date)
####################recoding the variables
patience_data$results <- ifelse(patience_data$results=="positive", 1,0)

####################grouping the data using date
pat_data <- patience_data %>% group_by(date) %>% summarise(results=n())
##############################################################################
####
tm<-slice(tm, 2:nrow(tm))

pat_data<-slice(pat_data, 2:nrow(pat_data))



pat<-tail(pat_data, -1)

final_data <- full_join(pat,tm,by="date")

final_data[is.na(final_data)] <- 0

final_data$results <- final_data$results.x + final_data$results.y

t_seriee_data <- final_data %>% select(date, results)

t_seriee_data$year <- format(as.Date(t_seriee_data$date, format="%d/%m/%Y"),"%Y")

plot.ts(t_seriee_data$results )
```

```
plot(t_seriee_data, type="o", ylab="covid cases",main="total covid cases recorded")


plot(tm, type="o", ylab="covid cases",main="covid cases recorded from workers")
plot(pat, type="o", ylab="covid cases",main="covid cases recorded from patients")


####################calculating the moving averages
library(forecast)
acf(tm, lag.max = 2, main="ACF PLOT FOR COVID19 CASE RECORDED FOR HEALTH WORKERS")
pacf(tm, lag.max = 2 , main="PACF PLOT FOR COVID19 CASE RECORDED FOR HEALTH WORKERS")
##############################################
acf(pat,  main="ACF PLOT FOR COVID19 CASE RECORDED FOR PATIENTS")
pacf(pat, lag.max = 2 , main="PACF PLOT FOR COVID19 CASE RECORDED FOR PATIENTS")
#####################################


library(fpp2)
library(forecast)
library(caTools)


sample <- sample.split(t_seriee_data$results, SplitRatio = 0.7)
train  <- subset(t_seriee_data, sample == TRUE)
test   <- subset(t_seriee_data, sample == FALSE)


ses.tserie <- ses(train$results, alpha = .2, h = 100)
autoplot(ses.tserie)


train.dif <- diff(train$results)
```

```
ses.train.dif <- ses(train.dif, alpha = .2, h = 100)
autoplot(ses.train.dif)


test.dif <- diff(test$results)
summarise(ses.train.dif)

# refit model with alpha = .05
ses.tr.opt <- ses(test$results, alpha = .2, h = 100)
autoplot(ses.train.opt)

# performance eval
summary(ses.train.opt)


# plotting results
p1 <- autoplot(ses.train.opt) +
  theme(legend.position = "bottom")
p2 <- autoplot(ses.train.opt) +
  autolayer(ses.train.opt, alpha = .2) +
  ggtitle("Predicted ")

gridExtra::grid.arrange(p1,p2, nrow = 1)
```