BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE AND ENGINEERING

DEPARMENT OF STATISTICS AND MATHEMATICS

MODELLING AND FORECASTING ZIMBABWE'S COVID 19 CONFIRMED CASES USING TIME SERIES ANALYSIS

BY
GWENHURE ANDY
B193400B

A DISSERTATION SUBMITTED TO BINDURA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE BACHELOR OF SCIENCE *HONOURS DEGREE IN STATISTICS AND FINANCIAL MATHEMATICS*

SURPERVISOR: MR MUKONOWESHURO

JUNE 2023

APPROVAL FORM

The undersigned certify that they have read and recommended to the Bindura University of Science Education for acceptance of a dissertation entitled "MODELLING AND FORECASTING ZIMBABWE'S COVID 19 CONFIRMED CASES USING TIME SERIES ANALYSIS" Submitted by GWENHURE ANDY, Registration Number B193400B in partial fulfillment of the requirements for the Bachelor of Science Honor's degree in Statistics and Financial Mathematics.

GWENHURE ANDY…………*a.gwenhure*…………….. …09 /06/2023…………………..

B193400B Signature Date

Certified by ………… ……………………………………………………

Mr. MUKONOWESHURO ………………………………………..

SUPERVISOR Signature Date

Certified by ………… ………………………………………………..

………………………………………………………………………………………………

MR M. MAGODORA ………………………………………

Chairman of Department Signature Date

………………………………………………………………………………………………

DEDICATION

I dedicate this dissertation to Masami Gwenhure, Getrude Bwanya, Abilly Gwenhure and Audry Gwenhure who have made sacrifices towards my personal and professional endeavors. I thank you for believing in my dreams.

ACKNOWLEDGES

My deepest appreciation goes to all those who made my work related learning a success .A special gratitude goes to the Bindura University of Science Education staff for the knowledge and skills fundamental statistical concepts they vested in me . I have to appreciate the guidance of my Supervisor Mr.Mukoweshuro.

I would like to express my sincere gratitude to lovely friends namely Godknows Murenza, Joshua Mukandawire, Alex Nyakwima and Tinashe Chiguvi for your exceptional mutual aid during my research project. Deprived of your persistent assistant, direction, inspiration, explanation, continues guidance, support, ideas and constructive criticism throughout the research project in was going to be unbearable.

Special thanks also goes to my father and mother for paying school fees and for being my pillar of strength, it would not be an easy expedition devoid of your support and reassurance.

ABSTRACT

The aim of this study is to forecast the Number of COVID-19 confirmed cases using SARIMA models from (2020 - 2023).The researcher focused on time series analysis of COVID-19 data with the main goal of assessing the trends, building a SARIMA model that will be used to forecast the Number of COVID-19 for the month of March 2023 to October 2023. The study applied the necessary descriptive research design and used the Box-Jenkins methodology in building Seasonal Autoregressive Integrated Moving Average ARIMA models. The R-Software was used to perform data analysis and (ARIMA) (0. 1. 1) (0. 1. 0) (12) model was implemented to predict monthly indexes for 8 points. The trend of Covid-19 cases in Zimbabwe up until now has been fluctuating. The number of cases has been increasing and decreasing, depending on various factors such as the public's adherence to safety guidelines, availability of vaccines, and the effectiveness of the government's mitigation strategies.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

R	A licensed free software programming language
AR	Autoregressive Models
MA	Moving Averages
ARMA	Mixed Autoregressive Moving Averages Models
ARIMA	Autoregressive Integrated Moving Average Models
SARIMA	Seasonal Autoregressive Integrated Moving Average Models
ACF	Autocorrelation Functions
PACF	Partial Autocorrelation Function
WHO	World Health Organization
MoHCC	Ministry of Health Child Care
COVID-19	Corona Virus @2019

CHAPTER ONE

1.0 Introduction

The research study's foundation is laid forth in this chapter. It looks to find out why the researcher hoped to use time series modeling and forecasting to model and predict Zimbabwe's COVID-19 confirmed case. The context in which the study was conducted, the issue that the research aimed to address, objectives of the research study, potential limitations encountered, and consequential significance for varied stakeholders who are interested in the research study's findings are the main topics of this chapter.

1.1 Background of the study

The initial detection of COVID-19 occurred in December 2019 in Wuhan, China. Humans can get upper respiratory illness from viruses like corona virus. A novel corona virus that causes severe respiratory symptoms in people was identified in 2019. The novel viral agent currently identified as the cause of the ongoing pandemic has been categorized as severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) owing to its striking resemblance to the previously known corona virus that precipitated severe acute respiratory syndrome (SARS) in 2003. The COVID-19 respiratory illness is caused by the pathogen SARS-CoV-2.

Since the WHO proclaimed a pandemic on March 11, 2020, novel corona virus disease has infected more than 160 million individuals worldwide.COVID-19, which is a distinct human corona virus disease, has been classified as the fifth pandemic since the 1918 flu pandemic.

The term "novel" denotes a virus novelty to human. Several COVID-19 vaccines have been authorized and are now accessible in many countries, thereby initiating significant immunization campaigns. Other precautions include maintaining a physical or social distance, isolating and ventilating indoor areas, covering coughs and sneezes, washing hands, and avoiding touching one's face with dirty hands.

Numerous countries responded to the outbreak of COVID-19 by implementing measures such as the closure of schools, businesses, and international borders as the means of preventing the spread of the virus. The prevailing indicators of COVID-19 are characterized by pyrexia, wearing, and a nonproductive cough. Other people feel discomfort diarrhea, a sore throat, nasal congestion, and runny nose.COVID-19 symptoms tend to develop slowly are often mild, with some individuals being asymptomatic. Around 80% of patients make a full recovery without needing medical treatment, while 1 in 6 COVID-19 patients experience severe respiratory problems. Individuals who are elderly or have preexisting health conditions like high blood pressure, heart disease, and diabetes are more susceptible to experiencing severe symptoms of COVID-19. Individuals who are experiencing symptoms such as fever, coughing, and difficulty breathing should seek medical assistance. Corona viruses transmission of the aforementioned can occur through interpersonal means, typically through close contact with an infected individual in settings such as workplace, home, and healthcare facilities.

First confirmed COVID-19 case in Zimbabwe was discovered in the month of March 2020, at which point the authorities declared a nationwide lockdown. The government imposes limitations such as halting schools and banning intercity travel during the lockdown. Zimbabwe adheres to COVID-19 guidelines, which include staying at home and keeping a social distance of at least one meter. Zimbabwe reported 56 verified cases two months after the initial case had been confirmed, including 27 ongoing cases, 4 deaths, and 25 recoveries. By the end of October 2020, during the initial wave of the outbreak in the country, which began with the first reported cases, the number of recorded cases exceeded 8000, 243 individuals also lost their lives due to COVID-19. Zimbabwe has experienced multiple waves of COVID-19, with the third wave peaking in July 2021, resulting in a significant increase in cases within just two months. Between November 2020 and January 2021, the peak of the second surge of the epidemic was attained, resulting in a total of 10034 cases and 960 fatalities. The fourth surge, which was primarily triggered by the Omicron variant demonstrated significant community transmission, as evidenced by a consistently high positive rate of approximately 35% on multiple days. The first and third waves occurred during the winter months, while the second and fourth waves coincided with the holiday season in December, which is known for increasing cross-border migration. As of January 29, 2022, Zimbabwe had recorded a total of 228948 cases, 5321 fatalities, 216028 recoveries, and 7594 active cases.

1.2 Statement of the problem

The COVI-19 pandemic has created a sense of uncertainty in various critical areas of global and national society. These include, but are not limited to, healthcare, the economy, travel restrictions, education, and social gatherings. The pandemic has brought about a significant impact on everyday life, with millions of people worldwide being affected. For instance, it is unknown how the shutdown of schools last spring affected student achievement.

1.3 Aim of the study

The aim of this investigation is to forecast the Zimbabwe's COVID 19 confirmed cases.

1.4 Research objectives

1. To comprehend the prevailing pattern and disposition by the COVID-19 confirmed cases in Zimbabwe

2. To come up with an accurate mathematical model for predicting future COVID-19 confirmed cases in Zimbabwe

1.5 Research questions

1. What trend is taken by confirmed cases?

2. Can COVID-19 confirmed cases be accurately predicted through a mathematical model?

1.6 Significance of the study

1. The government and Ministry of Health, Child and Care (MoHCC) of Zimbabwe will find this study valuable in respect of coming up with COVID-19 restrictions. Looking on trends health sector can easily predict which season has highest COVID-19 confirmed and make decision tightening COVID-19 restrictions.
2. The researcher will contribute to the existing literature on SARIMA model on modeling and forecasting COVID-19 confirmed cases in Zimbabwe by examining its applicability.

1.7 Academic relevance

1. Academics will benefit from the research findings as it will be a source of reference for other students who might want to carry out research on modeling and forecasting in the future.

1.8 Assumptions of the research study

This research study is bound by the subsequent assumptions:

1. Data of Zimbabwe's COVID 19 confirmed cases are the true representation of the events that occurred.

2. The data does not contain any significant outliers that could affect the model’s performance.
3. The data used for the research is accurate and reliable.

1.9 Limitations of the study

The following list of restrictions the researcher encountered while working on this project:

1. There was an electric power outage during the study, which caused editing and typing to be problematic.
2. There was a network connection issue that made it difficult to do research and access per tinent theoretical and empirical papers online.
3. Because of time constraints, the researcher was unable to examine alternative forecasting techniques.

1.10 Delimitation of the study

1. The study’s analysis does not account for additional variables that may impact the transmission of the virus, such as demographic characteristics, socioeconomic status and heath care infrastructure.
2. The research does not take into account the impact of new variants of the virus or changes in government and social distancing measures over time.

1.11 Summary

The study's guidelines are provided in this introductory chapter, which also explains what the study is about. This chapter served as an introduction to the entire study. This chapter furnishes an elaborate overview provided of the research questions, objectives, and problem statement. Next chapter will concentrate on a literature review of time series analysis, specifically regarding COVID-19 cases confirmed in Zimbabwe.

CHAPTER TWO: LITERATURE REVIEW

2.0 Introduction

Conceptual and theoretical foundation upon which the study is built is highlighted in Chapter 2. Based on what the pertinent theory and research finding on the subject show, it gives the framework for performing the research investigation. The chapter examines time series studies conducted by different academics on COVID-19 confirmed cases in Zimbabwe. The literature review highlights a difference between the researcher's interest and their level of expertise in the subject matter.

2.1 Theoretical framework

The initial of COVID-19 cases were reported in China in December 2019, according to Murewanhema et al. (2020). Since then, the illness has become a pandemic that affects the entire world. On March 20, 2020, Zimbabwe reported its first case, and since then, the number has been rapidly rising. From March 20th to June 27th in the year 2020, the Zimbabwe Ministry of Health and Child Care (MoHCC) collected and released daily reports on events that occurred during that timeframe. The daily situation reports did not impute missing data SARS-CoV-2, the virus known as COVID-19 elicits severe acute respiratory syndrome, according to Srivastava et al. (2020) and Guo et al. (2020). The announcement of a worldwide health emergency concerning COVID-19 was made by the World Health Organization (WHO) on April 16th, 2020, according to Remuzzi et al. (2020).

As of 1 July 2020, the World Health Organization (WHO) has designated the African continent as the fifth most affected region. As of 1 July 2020, South Africa held the top spot with 151209 cases and 2657 fatalities, followed by Egypt with 68311 cases and 2957 fatalities. The Government of Zimbabwe (2020) reports that on March 20, 2020, Zimbabwe confirmed its first COVID-19 case. Zimbabwe reported 56 verified cases two months after the initial case had been identified, including 27 ongoing cases, 4 fatalities, and 25 recoveries. The first phase of an epidemic in a country, which commences with the identification of the initial case, as of October 2020's conclusion, there were roughly 8000 cases of the disease, and 243 fatalities were attributed to COVID-19, according to Kavenga et al. (2021). Murewanhema et al. (2021) claim that there were a maximum of 960 fatalities during the second wave of the epidemic reached its highest point between November 30, 2020, and January 31, 2021, based on the overall number of cases, which were 10034 cases and a maximum of 960 fatalities. Murewanhema et al. (2021) claim that the third wave peaked in July 2021, causing an increase in COVID-19 cases of roughly 38000 to 120000 over the course of two months. The Omicron (B 1.1.529) variation was primarily responsible for the fourth wave which had a positivity rate of about 35% on numerous days, indicating widespread community transmission and perhaps indicating that the nation is undergoing testing. The first and third waves peaked in Zimbabwe during winter, whereas the second and fourth waves coincided with season in December, which is linked to increased mobility and cross-border migration, according to Madziva et al. (2021). The nation had 228948 cases and 5321 fatalities as of 29 January 2022, as well as 216028 recoveries and 7594 ongoing cases. Zimbabwe closed all face-to-face learning completely between March and mid-September 2020, reopened partially between mid-September and October, and then completely between November and December 2020. Due to school closures, there are now more online courses available. A small proportion of rural students (only 9%) and a relatively larger proportion of urban students (40%) reported using mobile apps for learning during school closures caused by the pandemic. This suggests that to access to learning resources was limited, particularly for students residing in rural areas. The majority of government schools have limited or restricted access to information and communication technology (ICT) resources, Zinyemba et al. (2021) found that the outcomes of online learning varied across government schools. In accordance with WHO, COVID-19 recommendations, the most esteemed educational institutions were capable of transitioning to a digitalized educational environment, which exacerbated the pre-existing divide between the affluent and underprivileged populations, furthering the inequality that existed before the COVID-19 pandemic, according to Hove et al. (2021).

2.2 Analysis of time series

According to Gregory (2008), time series comprises a set of data points that are measured repeatedly over time.

The process of time series analysis includes identification of three key components: trend, seasonal and irregular.

2.3 Components of Time Series

The three primary elements of time series analysis are trend, seasonality, and irregularity, with these components being identified and analyzed upon visualizing the time series data.

2.3.1 Trend

Montgomery (2008) a trend can be defined as the component of time series that depicts low-frequency fluctuations in addition to the removal of high and medium frequency oscillations. It is the observation of patterns that change through time, either in an upward or downward direction. In time series analysis, a trend refers to the long-term movement of the long-term movement of the data without any calendar-related or irregular effects, reflecting the underlying level of the series. It is the outcome of factors like population expansion, price increases, and broader economic shifts. A rising trend can be seen in the diagram below.

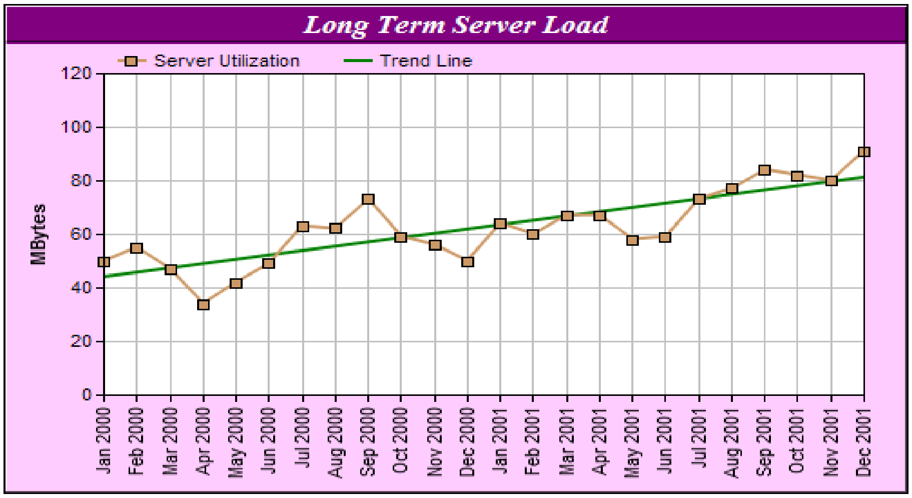


Figure 1: Showing an upward trend

2.3.2 Seasonality

Seasonality alludes to an ongoing cycle of occurrences. A pattern that, after some time, keep happening. The seasonal component of a time series is made up of effects that tend to be consistent in their timing, direction, and amplitude. These effects can be related to seasonal fluctuation or changes in demand that are tied to the calendar. It results from regular, calendar-related stimuli like the seasons' appropriate weather swings, business and administrative processes like the beginning and end of the school year, and cultural practices like Christmas. Below is a seasonality graph.

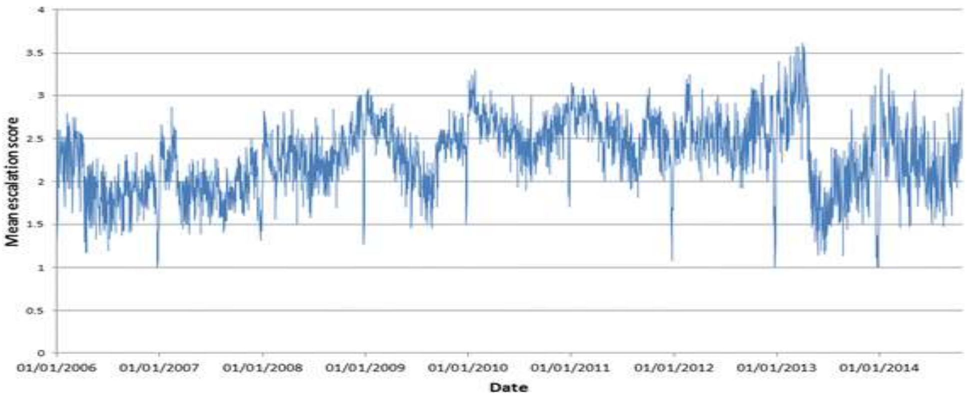


Figure 2: Seasonality graph

2.3.3 Irregular (unsystematic)

Occasionally referred to as the Residual. After estimating the seasonal and trend from a time series, the irregular component is what is left. Irregularities in a time series are unpredictable and inconsistent brief variation that can obscure the trend and seasonality of the data, potentially dominating movement in a highly erratic series. The irregular graph is shown in Figure 3 below.

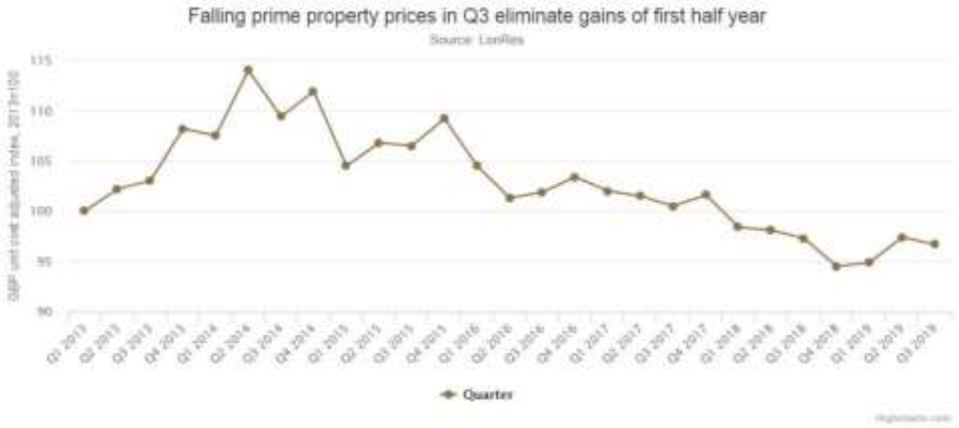


Figure 3: Showing an irregular trend

2.4 Models in Financial Time Series

2.4.1 Autoregressive (AR) Model

Lags, which are utilized by autoregressive models, refer to a time series forecast that relies entirely on the historical values of the series (Cryer, 1986). In Autoregressive (AR) Model, the Partial Autocorrelation Coefficient Function (PACF) had a sizable spike, and the Autocorrelation Coefficient Function (ACF) dropped sequentially, according to Da Hye et al. (2021). The number of significant spikes in the PACF is used to establish the order of the AR (p) Model. The stochastic difference equation for the AR (p) model process is given as follows.

$xt = \delta + \alpha 1xt{-}1 + \alpha 2xt{-}2 + \cdots + \alpha pxt{-}p + \mu t$ with $\alpha p \neq 0$, in this context, μt represents a completely random process that has an average of zero and a variance $\sigma 2$

2.4.2 Moving Averages (MA) Model

The moving average model makes predictions about a series using ‘error lags’, which are historical errors in the series. Da Hye et al (2021) report that there is a large surge in the ACF. The order q of the MA model is established based on the number of ACF significant spikes, and the PACF decreases sequentially. The next event will be the mean of the previous event according to this model, which takes into consideration very short-run autocorrelation of time series. Looking at the ACF plot of the time series can typically be used to estimate the moving average model’s order q. The MA (q) model’s formula is as follows:

$Xt = \varepsilon + \varepsilon t - \phi 1\ (\varepsilon t - 1) - \phi 2\ (\varepsilon t - 2) - \cdots - \phi q\ (\varepsilon t - q)$

With, $\phi q \neq 0$ and εt is the white noise process

2.4.3 Autoregressive Integrated Moving Averages (ARMA) Model

The ARMA model is created by merging the AR and MA models. If the equation for the first-order AR model reaches the initial point, it will result in an infinite moving average. In order to apply the ARMA model, it is required to specify the values of p and q. The value of corresponds of important terms in the autocorrelation function (ACF), while the value q refers to the number of significant terms in the partial autocorrelation function (PACF). If a time series has an ARMA (p, q) model, it is said to exist.

$yt = \delta + \emptyset 1yt{-}1 + \emptyset 2yt{-}2 + \cdots + \emptyset pyt{-}p + \varepsilon t - \theta 1\varepsilon t{-}1 - \theta 2\varepsilon t{-}2 - \cdots - \theta q\varepsilon t{-}q$, where εt is a white noise process.

2.4.4 Autoregressive Integrated Moving Averages (ARIMA) Model

Prior comprehension of the ARMA model is crucial in comprehending the ARIMA model. The ARIMA model is represented by the notation ARIMA (p, d, q), where p and q have the same meaning as in the ARMA model. However, d specifies the number of initial differences (Yu and Zhang, 2004). A development of the autoregressive (AR) and moving average processes (MA) as special examples, according to William, Wei (2006), is an ARIMA model. In order to verify stationarity, the Augmented Dickey-Fuller (ADF) test is used. Musundi et al. (2016) explain that the Box-Jenkins approach compromises four phases: identification, estimation, diagnostic checking, and forecasting of time series, and it is an iterative process. The roots of the associated polynomial are used to diagnose the ARMA and AR components because of their identical stationarity in the modeling process. Because the is coefficients in its MA representation do not eventually decrease to zero, an ARIMA model , like a random walk model, has great memory because the past shock at model point 1 has a lasting impact on the series. Utilizing differencing is a common strategy for managing unit-root non-statitionarity. ARMA (p, q) is stationary if all 18 of the equation's roots have absolute values that are less than one. The exponential decline of the ACF and PACF of an ARMA (p, q) can be demonstrated (Tsay 2010). This is how the generic model might be stated.

$\boldsymbol{\phi\ (B)\ Zt = \Theta\ (B)\ a\mathrm{t}}$

In the model for COVID-19 confirmed cases, $\boldsymbol{Zt}$ is the variable being modeled, the unknown model parameter ϕ and Θ for a white noise process are estimated using the method of estimation which either least-squares or maximum likelihood, and the backward difference operator $\boldsymbol{B}$ is also employed.

2.4.5 Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

In order to accommodate seasonality in the data, SARIMA models were introduced. The SARIMA model, proposed by Box and Jenkins, is appropriate for analyzing both seasonal and non-stationary data. Although seasonality is distinct from stationarity, seasonal relationship cannot be completely avoided but can only be partially managed by stationarity. The SARIMA models are represented by the mathematical formula SARIMA (p, d, q) (P, D, Q) s. Here, P signifies the count of seasonal AR terms, D indicates the number of seasonal differences, Q represents the count of seasonal MA terms, and s stands for the length of the season. The model will perform better if seasonality is eliminated, but eliminating seasonality in data is a challenging undertaking. This is how a generic model might be stated.

$$\boldsymbol{\phi_{\mathrm{p}}\ \Phi_{P\mathrm{s}}\ (B^{s})\ \nabla^{d}\nabla^{D\mathrm{s}}Z_{t} = \theta q\Theta_{Q\mathrm{s}}\ (B^{s})\ a_{\mathrm{t}}}$$

In a seasonal time series model, the seasonal lag is denoted by s, while $\boldsymbol{B}$ refers to the shift operator in the backward direction, $\boldsymbol{B^{s}}$ represents the seasonal backshift operator, $\boldsymbol{\nabla^{d}}$ represents the operator for differencing, and $\boldsymbol{\nabla^{Ds}}$ represents the operator for seasonal differencing.

2.4.6 Autocorrelation Functions (ACF)

The autocorrelation function indicates that the correlation between signal values changes as the interval between them varies. The autocorrelation function quantifies the persistence of a stochastic process in the time domain and does not provide any details about the frequency characteristics of the process. In order to choose an appropriate ARIMA model, autocorrelation function (ACF) is helpful (Cuhadar 2014).

2.4.7 Partial Autocorrelation Function (PACF)

When we assume that we are aware of and taking into consideration the values of another set of variables, we can use the partial autocorrelation function to determine the correlation between two variables. A stationary time series' partial correlation is calculated using the partial autocorrelation function, with the time series' lag value being regressed at all lesser delays. In order to choose an appropriate ARIMA model, PACF is also helpful (Cuhadar 2014)

2.4.8 Information Criteria and Akaike Information Criterion (AIC)

According to Richard and McElreath (2016), the Akaike information criterion (AIC) is an indicator of prediction error and evaluates the comparative effectiveness of statistical models for a particular dataset. AIC (Akaike Information Criterion) calculated the quality of statistical models in relation to other models, and can be used as a model selection method. It is rare for a model to perfectly depict the data generation process. The AIC assesses the amount of information lost when the model represents the data, with lower loss indicating a higher-quality model.

2.4.9 Schwarz-Bayesian Information Function (BIC)

The Bayesian information criterion is a technique used to choose a model from a limited set of models, also known as the Schwarz information criterion. It is preferred to use the model with the lowest BIC. It is crucial to note that the BIC can only be employed to compare estimated models if the numerical values of the dependent variable are accessible. Formally speaking, BIC is:

BIC=ķ Ln (n) -2Ln (′L)

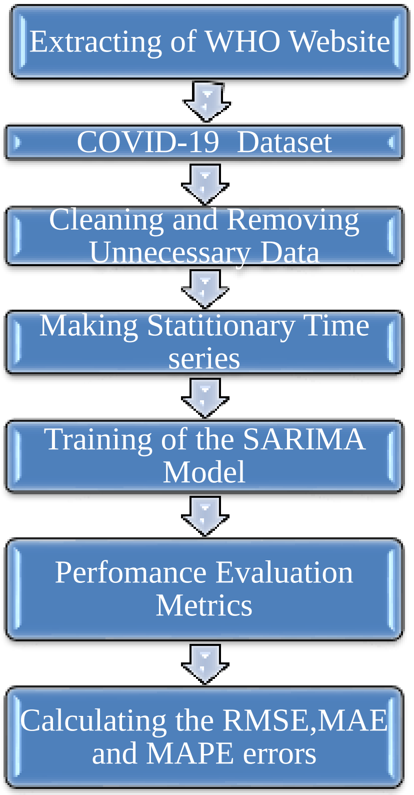
Let ′L be the maximum likelihood estimator of the model **M,** such that **L =p(x/α̉̀, M**). When fitting a statistical model to data, the values of the parameter that maximize the likelihood function are represented by **α̉̀.** K stands for the count of parameter estimated by the model, while n represents the number of data points in the variable x.

2.5 Model Checking

In this study, the suitability of the model is assessed using several metrics. Several metrics, such as MAPE, MAE, RMSE, and MASE, can be employed to assess the precision of the model and compare the effectiveness of various models. Statistics uses the term mean absolute percentage error (MAPE) to describe how well a forecasting system makes predictions. The accuracy is typically expressed as a ratio. The mean absolute error (MAE) is a measure used to evaluate the discrepancies between corresponding observations that describe the same phenomenon. RMSE is another metric that measures standard deviation of the residuals or prediction mistakes. Both metrics are commonly used in statistical modeling and prediction tasks to evaluate the precision of the models. The residuals show the degree of the deviation of the data points from the regression line. Root mean square error is a commonly employed metric in climatology, forecasting, and regression analysis to validate experimental findings. The issues present in the other measurement are absent from the widely applicable Mean Absolute Scaled Error (MASE) assessment of forecast accuracy. J. Rob Hyndman 2006.

2.6 Conceptual framework

The important elements, variables, and relationships pertinent to a specific phenomenon or research issue are identified and arranged in the conceptual framework, which is a theoretical framework or model. It offers a methodical approach to thinking through and evaluating complicated events and aids in forming research topics and study plans. It is possible to utilize a conceptual framework to describe how a certain intervention or occurrence affected the results.


Extracting of WHO Website
COVID-19 Dataset
Cleaning and Removing Unnecessary Data
Making Statitionary Time series
Training of the SARIMA Model
Perfomance Evaluation Metrics
Calculating the RMSE,MAE and MAPE errors

2.7 Empirical evidence

In Korea, Da Hye et al. (2021) conducted a study with the goal of predicting COVID-19 confirmed cases using empirical data analysis. The aim of this identify was to determine an appropriate time point for making precise predictions about the COVID-19 cases in Korea. Utilizing the most recent data being processed for the third wave, they fitted an ARIMA model. They anticipated reaching 70000 total confirmed cases by the start of the next year (2022).

Makoni and Chikobvu (2018) used time series modeling and forecasting to model and project visitor arrival in Zimbabwe the Victoria Falls Rainforest as a case study. They create a time series diagram using raw data to show the behavior of arriving tourists. The ACF, PACF, and EACF plots all pointed to a SARIMA (2, 1, 0) (2, 0, 0)12 model that closely matches data on tourist arrivals. SARIMA (2, 1, 0) (2, 0, 0)12 model was found to be a better choice than the commonly used SARIMA (1, 1, 0) (1, 1, 0)12, as well as other models that were considered. This conclusion was based on a comparison of several metrics, including AIC, BIC, and forecasting accuracy measures. For the year 2018, they predicted a favorable increasing trend.

In a study on Analyzing Univariate Time Series for Short-term Forecasting using Artificial Neural Networks: A Case Study on Public Ambulance Emergency Preparedness by Mapuwei et al, 2020 used seasonal ARIMA models and two neural network models. Applying the Box-Jenkins methodology, they discovered that ARIMA models may be utilized as they provide directional with linearity (trend) across time, and ultimately they were able to achieve the best-fitted model as ARIMA (0, 1, 1) (0, 0, 2)12.

Perone (2022) conducted a study with the goal of applying SARIMA model to forecast the fourth global wave of cumulative COVID-19 deaths: findings from twelve severely affected major countries. Total number of deaths predicted by the study was to occur between August 21, 2021, and September 19, 2021. The results showed that: it is fantastic news that the implemented forecasting procedures showed strong accuracy of the forecast in both the test dataset. This was accomplished by outperforming less sophisticated alternate methods such as the Mean, Naïve, and Seasonal Naïve models. Furthermore, according to the AIC and almost all of the forecasting accuracy metrics, including MAE, MAPE, MASE, and RMSE, and SARIMA models demonstrated superior performance compared to the ARIMA models when predicting future values. This finding suggests that the time series data had strong seasonal patterns. The investigation also found that the SARIMA model's 30-day projection aligned well the observed data during the analyzed period of August 21, 2021, in almost all of the examined countries. As such, it can be concluded that SARIMA models are highly precise and dependant techniques for forecasting the total number of deaths caused by COVID-19.

Zhao et al. (2022) investigated ARIMA, GM (1, 1), for the prediction of TB cases in China, and LSTM models. They discover that the LSTM model, which performs more accurately than the ARIMA and GM (1, 1) models, was the best model. Their predictions' outcomes may serve as a guide for TB prevention strategies in Mainland, China.

2.8 Knowledge Gap

There haven't been many studies done yet that predict the COVID-19 cases in Zimbabwe. In addition to implementing the monitoring government policies on enhancing the eradication of COVID-19 in the, the aim of this investigation is to shed light on the health sector and potential consequences of the COVID-19 outbreak in Zimbabwe. Additionally, predicting future COVID-19 confirmed case numbers is significant measure of the nation's health.

2.9 Definition of Keywords

1. Time series analysis

Time series analysis, according to Tsay (2006), is an introduction to some statistical tools useful for analyzing series and gaining experience with various econometric methodologies' financial applications.

2. COVID-19

It is a contagious respiratory disease that is induced by the SARS-CoV-2 virus, which is a new strain.

3. COVID-19 CONFIRMED CASES

Number of individuals who have undergone testing and received a COVID-19 virus diagnosis.

2.10 Summary

The literature review at the top makes it obvious that time series models have been successful at forecasting the COVID-19 confirmed cases in Zimbabwe. However, a recorded quantitative data sample of the same historical sequence is required. These methods are more accurate for short-term forecasts (two to three years) because their precision declines with time. To get or produce more precise long-term projections, greater in-depth understanding of the variables of interest, a well defined model, and consideration of statistics like root mean square error, mean absolute percentage error, and many more are needed. Additionally, it's crucial to keep an eye on the forecast that were calculated and take them into account for any necessary modifications in the future.

CHAPTER THREE: RESEARCH METHODOLOGY

3.0 Introduction

Research methodology is a crucial aspect of any research study, as it outlines the steps and procedures that a researcher will use that a researcher will use to conduct their investigation. It provides a clear plan to keep researchers on track and ensure that the study is efficient, effective, and manageable. A well-designed research methodology is essential to ensuring that the study yields valid and reliable results that can be used to answer the research question and contribute to the body of knowledge in the field. In the following sections, we will examine various aspects of research methodology, including the research design, data collection techniques, and data analysis methods. This chapter examines the methods the researcher employed to collect information from the respondents. The researcher described the method for gathering the data in this chapter. The following section discusses the research strategy, research tools, data collection framework, data presentations, and analysis criteria used to produce the findings, interpretations, and conclusions.

3.1 Research Design

The research design refers to the plan for gathering, measuring, and analyzing data in a study. It involves selecting an overarching strategy to integrate all the different components of the research in a cohesive and logical way, with the aim of effectively addressing the research problem at hand. I employed a cohort design in this research project. A cohort study is a type of study that is typically conducted over a period of time with participants from the same populations the topic or representative participant and who share specific characteristics. The study used a quantitative research approach. When statistical findings are needed to gather crucial actionable knowledge, quantitative research is used. When making decisions, the understanding gained from complicated numerical data and analysis is quite beneficial.

3.2 Research instruments

There are many fact-finding strategies and tools for data collection, and the specific ones you choose will depend on the type of information you are trying to gather, as well as the resources available to you. The researcher used quantitative research instruments such Microsoft package excel to input and view data and R-Programming 4.3.0 for the analyzation and creating graphs of collected data.

3.3 Data source

Data source refers to the primary source of information that provides the necessary data for analysis and interpretation. The researcher used the data which were published by World Health Organization (WHO).

3.4 Data collection

The World Health Organization (WHO) department provides the researcher with information on the confirmed COVID-19 cases in Zimbabwe. It is important to verify the credibility and reliability of sources, especially when conducting research or analysis. If the information was compiled and put together by other organizations, it may be necessary to look at the original sources of the information to ensure accuracy and to access the credibility of the organizations that compiled the information. Additionally, it is important to consider any potential biases or perspectives that may have influenced the compilation of the information. Since it is challenging to find primary data on this topic, the researcher chooses to use secondary data. The data was acquired during a three –year period, beginning in March 2020 and concluding in February 2023.

3.4.1 Secondary Data

It sometimes goes by the name "second-party data", any dataset gathered by someone other than the person using it is referred to .The researcher utilized the internet to retrieve secondary data, which refers to information that has already been collected and analyzed by others for a purpose unrelated to the current matter at hand. It is essential to evaluate the accuracy and relevance of secondary data before including it in one's research or analysis. Such information is already available, and it can be obtained from both internal and external secondary sources, depending on the organization. Identifying primary versus secondary data is the issue at hand. Internal secondary data sources include things like old market research data, your company's precious financial records, and old sale reports. Reports created by external data sources are an example of an external data source. Government agencies, nonprofit organizations, business associations, trade associations, industry organizations, educational institutions, and private businesses are often used sources of secondary data. By accessing considerably larger and more varied samples than you might gather alone, using secondary data helps broaden the scope of your research. It also implies that the researcher has no say in the variable to be measured.

3.4.2 Advantages of Secondary Data

1. It is economical. Most of the secondary sources are available for free or at very low costs. Secondary research enables the collection of data without any financial investment. It saves the researcher's effort.
2. It is time saving which means secondary research can be conducted quickly. In many cases, it only takes a few Google searches to locate a data source for secondary research.
3. The use of secondary data aids in making primary data collection more precise since it enables the identification of gaps and deficiencies, as well as the determination of what additional information needs to be gathered.
4. Secondary data is very easy to access to anyone. Anyone can collect the data.
5. Secondary data is simpler to obtain than primary data and can be adapted to address various issues.

3.4.3 Disadvantages of Secondary Data

1. Secondary data can often be subject to bias, as it is sourced from external parties with their own agendas and perspectives.
2. Secondary data may not be entirely current, having been gathered in the past and potentially out of date.
3. Since the researcher does not have control over the collection and processing of secondary data, there is a risk that the data may be of poor quality or accuracy.
4. Some drawbacks of secondary data include that they may not be problem-specific, that they may be out-of-date and therefore inappropriate, that it may be challenging to judge their accuracy, that they may not be subject to further manipulation, and that combining different sources may result in inaccurate results.

3.5 Data cleaning

Not all times will the data be in a format that is ideal for developing models. Data cleaning seeks to organize unorganized data. A tidy dataset, according to Wickham (2014), can be characterized by three characteristics:

1. Each observation comprises a row,
2. a variable from the column and
3. An observational unit produces a table.

The data are shown in a tabular format and cover the entire period from 01 March 2020 to 28 06 February 2023. After the data has been pre-processed, each observation is represented by a row and each variable by a column. The monthly number of COVID-19 confirmed cases and the time in months are the two variables in the dataset. The accuracy and presence of duplicates in the data were verified.

3.6 Model identification

Data must be steady in order to fit a time series model. For stationary data, the lag k is the only factor that affects the mean, variance, and covariance, which are all constant. Before deciding on the model, we must first determine whether the data are stationary. This will be done using the Augmented Dickey-Fuller (ADF) test. If the data are not stationary, the researcher will use the differencing method until the data are. When the data are stable, the researcher can identify the model using the ACF and PACF graphs. The software will automatically choose the best model when the command auto.arima is given. The unit root test, minimization of Akaike Information Criterion with a correlation (AICc), and maximum likelihood function are then combined to uncover the ARIMA model, which are Hyndman and 25 Khandakar.

3.7 Diagnostic checking

The data's suitability for time series analysis will be evaluated using the model adequacy of both AR and MA models. To examine the behavior of the data, ACF and PACF will also be added. Montgomery et al. (2016) assert that the residuals ought to behave like a white noise process. Since there are no trends visible, a rectangular form indicates that the model is acceptable. If the model is acceptable, there should be no discernible structure in the residual sample autocorrelation function. The researcher will make use ARIMA Chart, which displays the methods employed to analyze the data using R Studio version 4.3.0.

3.8 Summary

The research design, methodology, and data collection process were all covered in this chapter. The chapter continued by outlining the data analysis steps the researcher will do in the following chapter. The analysis and display of data are covered in the following chapter.

CHAPTER FOUR: DATA ANALYSIS, PRESENTATION AND DISCUSSION

4.0 Introduction

This chapter focuses on data analysis process and also the interpretation of the research results. This study's main objective is to forecast Zimbabwe's COVID-19 confirmed cases. Therefore, it discusses the modeling of the ARIMA model being highlighted in brief. Moreover, data analysis and results are based on the time series data on this COVID-19 collected and calculated on monthly. It tries to point out the accurate model for this scenario, spotting the trend as well as predicting future values from the model.

4.1 Importing and Loading Data in *R-Software*

COVID-19 confirmed cases data was originally offered on the Excel (**xlsx**) sheets format. The researcher converted them from **xlsx** to **.csv** format for proper importing and loading of time series data. Data was converted into time series horizontally, with months (Jan-Dec) corresponding to their years from 2020-2023 before it was uploaded in R-Studio. Data cleaning checking for missing data was run and found no missing data. The descriptive statistics were calculated by the software by a coding summary (Data Frame) and obtained the output below:

TABLE 1: Data Summary in Excel

Minimum	8
Maximum	75691
Range	75683
Sum	264270
1st Quartile	787.50
Median	2055
3rd Quartile	5594
Mean	7340.83
Sample Variance	2.35E+08
Standard Deviation	15338.51
Kurtosis	13.3555
Skewness	3.574992
Count	36

The overall numbers of observations were 36 from March 2020 up to February 2023. Table 1 above shows a minimum COVID-19 confirmed cases of 8 and also the maximum COVID-19 confirmed cases of 75691. Moreover, the averages of the COVID-19 confirmed cases for this period were 7340.83, lower quartile 787.50, and the upper quartile of 5594.It also shows that the data skewed of about 3.574992.It shows that the total number of Covid-19 confirmed cases were 264270.

4.2 Data Visualization in Time Series

Data visualization presents information through visual aids such as time series graph, making it easier for people to understand the meaning of the data. This improves comprehension of large data sets by highlighting visible patterns, trends, and outliers.

CODE 1 in R Studio

```
> a1 <- read.csv ("C: /Users/TOBANYWAY/Desktop/ANDY PROJECT/a1.csv")
> View (a1)
> attach (a1)
> library (tseries)
> library (forecast)
> plot.ts (NUMBER.OF.COVID.19.IN.ZIMBABWE)
```

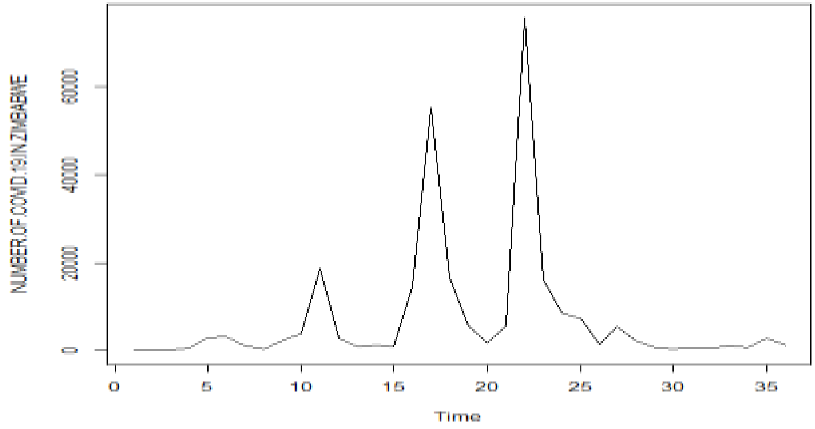


Figure 4: A Time series plot of Number of COVID-19 confirmed cases in Zimbabwe

Figure 4 presents a time series graph that illustrates the number of COVID-19 confirmed cases from March 2020 to February 2023. It can be noted that the plot in figure shows that the Number of C0VID-19 confirmed cases at first was found with slight fluctuation and it suddenly rose to reach approximately 18780 COVID-19 confirmed cases. The number of COVID-19 cases decreases and returns to its normal rate. Furthermore, there is an upward trend of COVID-19 confirmed cases in Zimbabwe which reach to maximum of 75691 the Number of COVID-19 confirmed cases showing a greater number of affected in the country and it decline. A trend promises a downward changes in the Number of COVID-19 confirmed cases. The graph above shows that COVID-19 confirmed cases is not stationary. This can be supported by the use of the Augmented DICKEY Fuller below.

4.3 Time Series Test for Stationarity (ADF Root Test)

This is the essential assumption test required when dealing with ARIMA model. In this case, the researcher should investigate this assumption if the time series data is stationary before venturing into any model building procedures. This study applies difference method as well as Augmented Dickey- Fuller test for root diagnosing.

The following hypothesis was tested as below:

H_0: The data is non-stationary

H_1: The data is stationary

The researcher executed a code in R software:

CODE 2 in R Studio

```
> adf.test (NUMBER.OF.COVID.19.IN.ZIMBABWE)
```

TABLE 2: Augmented Dickey-Fuller Test

AUGMENTED DICKEY-FULLER TEST
Data: NUMBER OF COVID-19 IN ZIMBABWE
Dickey-Fuller = -2.499
Lag order = 3
p-Value = 0.379
Alternative Hypothesis: Stationary

Table 2 shows the Dickey-Fuller value of -2.499, which is a statistical test used to ascertain whether a time series is stationary or not-stationary. Table 2 display the quantity of lagged differenced applied in the test, as well as the p-value, which is the probability of observing a value as extreme as -2.499. The p-value of 0.379 is greater than the conventional threshold of 0.05, indicating that we cannot reject the null hypothesis of non-stationary at a 5% significance level.

CODE 3 in R Studio

```
> A = diff (log (NUMBER.OF.COVID.19.IN.ZIMBABWE))
> plot.ts (A)
```

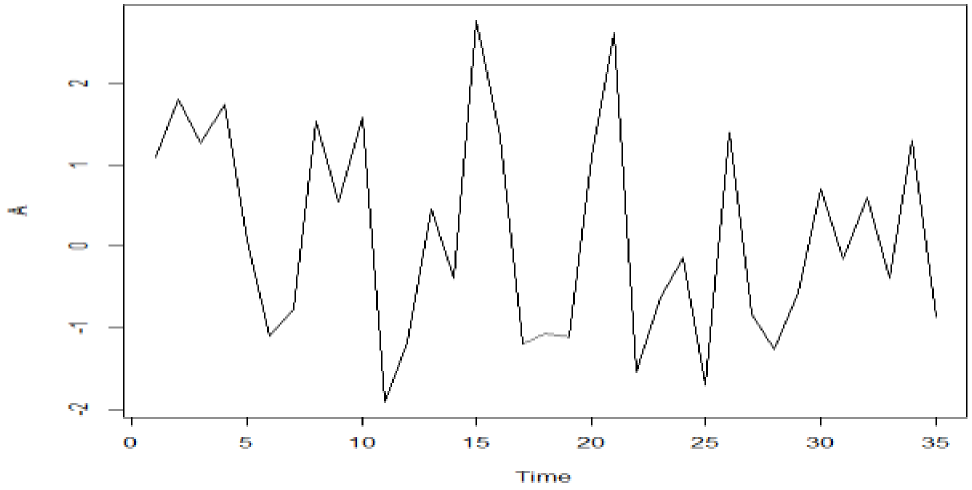


Figure 5: Differenced plot for Number of COVID-19 in Zimbabwe

Figure 5 above shows that our data is now converted into stationary data. This can be further proved by the Augmented Dickey Fuller below.

CODE 4 in R Studio

```
> adf.test (A)
```

TABLE 3: Augmented Dickey-Fuller for Differenced Data

AUGMENTED DICKEY-FULLER TEST
Data: A
Dickey-Fuller = -6.3923
Lag order = 3
p-Value = 0.01

Alternative Hypothesis: Stationary

Table 3 show the results of an Augmented Dickey-Fuller test performed on a variable called ‘A’. ‘A’ represent differenced Number of COVID-19 cases in Zimbabwe. The Dickey-Fuller value is -6.3923, which indicates that the variable is stationary. The lag order of 3 refers to the number of lags used in the regression equation. With a p-value of 0.01, there is a substantial evidence to reject the null hypothesis that the variable is non-stationary. The low p-value provides evidence for the alternative hypothesis that the variable is stationary. Overall, the results suggest that the variable has a stable mean and variance over time and is not affected by any long-term trends or seasonal effects.

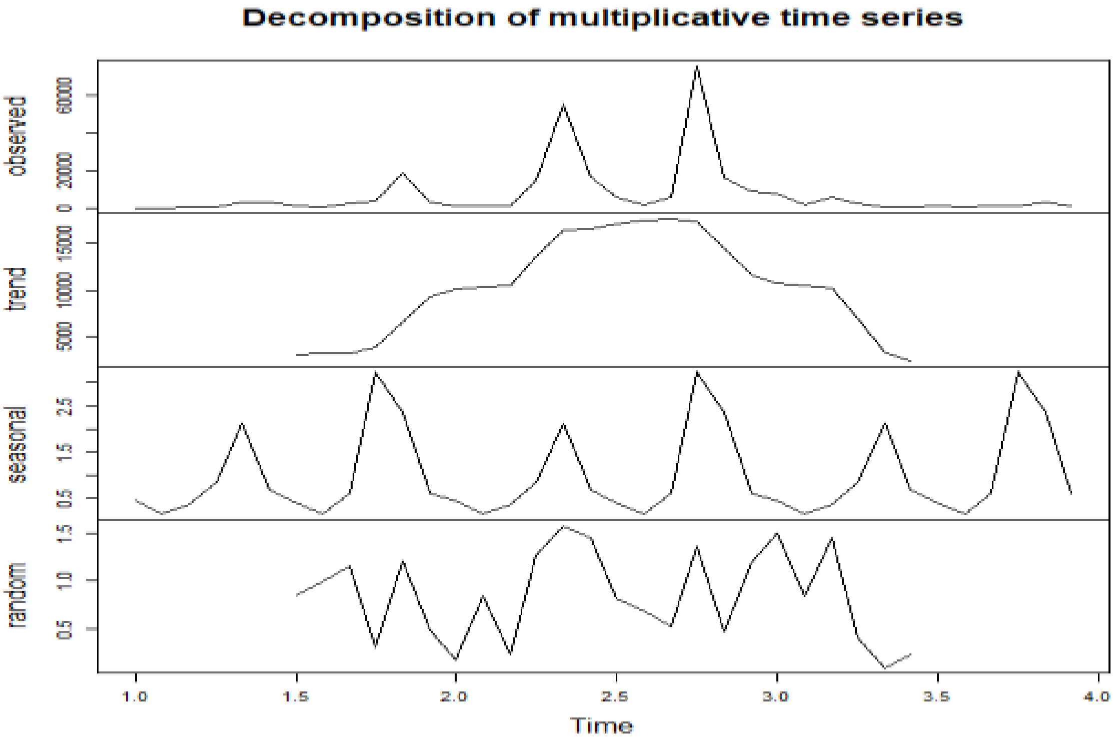
4.4 Visualization Time Series Components Individually

Figure 6 shows a decomposed multiplicative time series and visualization of the following three time series components which are seasonal, trend and random as indicated below:

CODE 5 in R Studio

```
> library (ggplot2)
> tsdata = ts (NUMBER.OF.COVID.19.IN.ZIMBABWE, frequency = 12)
> ddata = decompose (tsdata,"multiplicative")
> plot (ddata, title (main = "Decomposition of the Data into Time-Series Components"))
```

Figure 6: Decomposition of Multiplicative Time Series



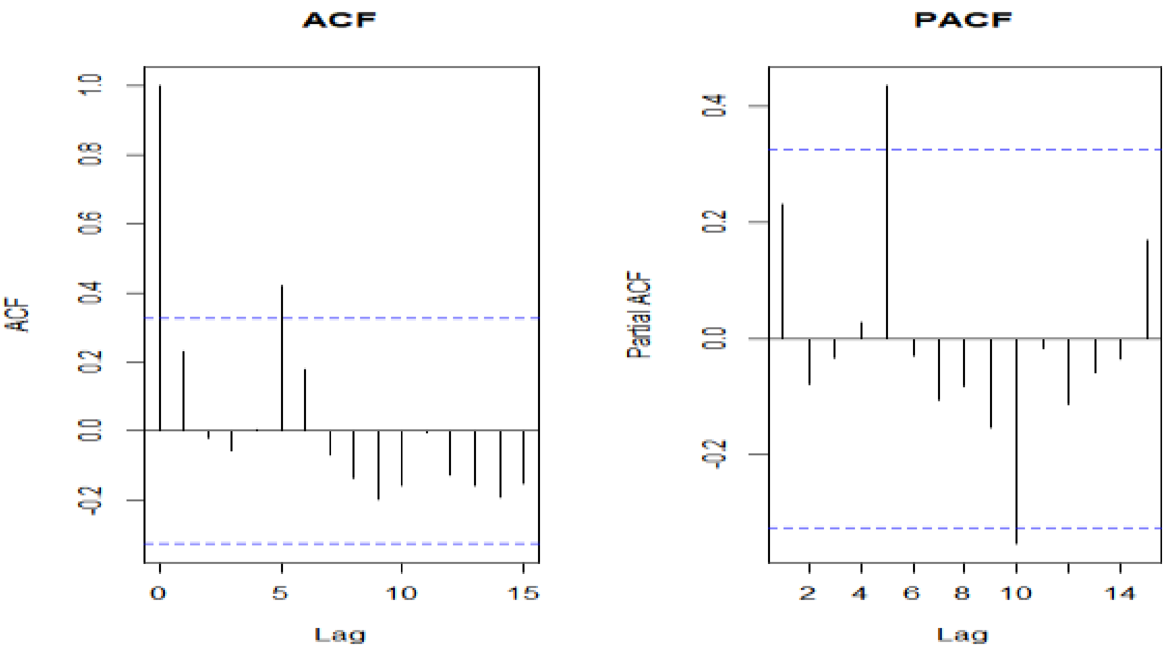
4.5 Testing for Autocorrelation (PACF) and (ACF) Index

The ACF and PACF plots portrayed above reflect about how observations in time series are being indexed over time and how well are they related to each other. The two plots clarify the order concept in Autoregressive and Moving Averages in time series analysis.

CODE 6 in R Studio

```
> par (mfrow = c (1, 2))
> acf (NUMBER.OF.COVID.19.IN.ZIMBABWE, main = "ACF")
> pacf (NUMBER.OF.COVID.19.IN.ZIMBABWE, main = "PACF")
```

Figure 7: ACF and PACF for Raw Data



4.6 Model Identification

At this stage, the researcher will identify the best fit time series model for Number of Covid-19 confirmed cases in Zimbabwe. The researcher will identify the Auto-Regressive and Moving Average terms which suits the data. The following command will be used:

The results obtained after running the command showed that ARIMA (0.1.1) (0.1.0) [12] is the best model for the data as shown in table 5 below.

Referring to the above Figure 7, it can be observed that the PACF does not cut off, indicating that p=0. Additionally, the ACF cuts off after lag 1, suggesting that q=1. These values are based on the assumption that we conducted the first difference on the data. Therefore, the model obtained becomes ARIMA (0, 1, 1) with the seasonality of order (0, 1, 0) that repeats itself in December per year that is after 12 months. There is no seasonal Auto-Regressive component (P=0), one seasonal first-order difference (D=1) and no seasonal Moving Average component (Q=0). [12] Specifies the seasonal of the model. In this case, the data is assumed to have a seasonal pattern with a period of 12 months that is monthly data.

Therefore ,the ARIMA(0.1.1)(0.1.0)[12] model is a time series model that includes a first-order difference and a moving average term for the non-seasonal component and a first-order seasonal difference for the seasonal component, with a seasonal period of 12 months. The best model was determined by the researcher through the calculation of the maximum likelihood function using AI

C and BIC functions. The analysis revealed that the optimal model is ARIMA (0, 1, 1) (0, 1, 0) [12].

CODE 7 in R Studio

```
> library (ggplot2)
> COVID = na.locf (NUMBER.OF.COVID.19.IN.ZIMBABWE, fromLast = TRUE)
> COVID = ts (COVID, start = 2020, frequency = 12)
> COVID
```

TABLE 4: Time Series of Number of COVID-19 cases in Zimbabwe

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2020			8	24	146	520	2961	3178	1048	482	2250	3874
2021	18780	2787	853	1349	908	14497	55195	16565	5669	1860	5569	75691
2022	16188	8617	7462	1361	5496	2372	668	377	762	652	1181	796
2023	2908	1216										

```
>auto.arima (COVID, D=1, ic = “aic”, trace = TRUE)
```

TABLE 5: Modeling SARIMA

ARIMA(2.1.2)(1.1.1)[12]	Inf
ARIMA(0.1.0)(0.1.0)[12]	546.818
ARIMA(1.1.0)(1.1.0)[12]	Inf
ARIMA(0.1.1)(0.1.1)[12]	Inf
ARIMA(0.1.0)(1.1.0)[12]	Inf
ARIMA(0.1.0)(1.1.1)[12]	Inf
ARIMA(1.1.0)(0.1.0)[12]	544.8943
ARIMA(1.1.0)(0.1.1)[12]	Inf
ARIMA(1.1.0)(1.1.1)[12]	Inf
ARIMA(2.1.0)(0.1.0)[12]	544.0708
ARIMA(2.1.0)(1.1.0)[12]	Inf
ARIMA(2.1.0)(1.1.0)[12]	Inf
ARIMA(2.1.0)(1.1.1)[12]	Inf
ARIMA(3.1.0)(0.1.0)[12]	544.4576
ARIMA(2.1.1)(0.1.0)[12]	542.5244
ARIMA(2.1.1)(0.1.1)[12]	Inf
ARIMA(2.1.1)(1.1.1)[12]	Inf
ARIMA(1.1.1)(0.1.0)[12]	541.1966
ARIMA(1.1.1)(1.1.0)[12]	Inf
ARIMA(1.1.1)(0.1.1)[12]	Inf
ARIMA(1.1.1)(1.1.1)[12]	Inf
ARIMA(0.1.1)(0.1.0)[12]	539.2184
ARIMA(0.1.1)(1.1.0)[12]	Inf
ARIMA(0.1.1)(1.1.1)[12]	Inf
ARIMA(0.1.2)(0.1.0)[12]	541.1839
ARIMA(1.1.2)(0.1.0)[12]	Inf
Best Model	ARIMA(0.1.1)(0.1.0)[12]

CODE 8 in R Studio

```
> model = auto.arima (COVID, D=1)
> summary (model)
```

TABLE 6: Summary of the Model

Series: COVID

SARIMA (0.1.1) (0.1.0) [12]

Coefficients:

Ma 1	-0.7575
s.e	0.1301
Sigma^2 = 753210228	log likelihood = -267.61
AIC	539.22
AICc	539.82
BIC	541.49

Training Set Error Measures:

	Training Set
ME	-2460.809
RMSE	21454.49
MAE	11649.91
MPE	-491.5673
MAPE	734.8522
MASE	0.7511732
ACF1	-0.01788781

The summary provides the results of a seasonal ARIMA (SARIMA) model fit to a time series data named ‘COVID’. The model has parameters (0, 1, 1) (0, 1, 0) [12] that represent non-seasonal difference, seasonal difference, and seasonal periodicity, respectively. The model coefficient values are also listed, with ma1 being -0.7575, and its standard error being 0.1301. The variance of the error is estimated to be 753210228, and the log-likelihood is -267.61. The information criteria such as AIC (539.22), AICc (539.82), and BIC (541.49) are presented, which help in evaluating the model’s goodness of fit.

The ‘Training Set Error Measures’ selection gives the performance of the model on the training set, for accuracy (ME, RMSE, MAE), percentage errors (MPE, MAPE), and statistics (MASE, ACF1). These metrics can be used to evaluate how well the model has fit to the training data. For example, the MASE value of 0.7511732 is less than 1, indicating that the model that the model provides an accurate forecast compared to naïve model that always forecasts the mean. However, the MAPE of 734.8522 indicates high forecasting errors on average.

4.7 Diagnostic Checking

At this point, the researcher will check if all the assumptions of ARIMA model are fulfilled. That is, stationarity, normality and Independence. Below is the command used by the researcher to run data in R studio for checking stationarity:

CODE 9 in R Studio

```
> plot.ts (model$residuals, main = "Model Residuals")
```

The graph in figure below clearly shows a white noise structure whereby the residuals deviated around mean zero and a constant variation.

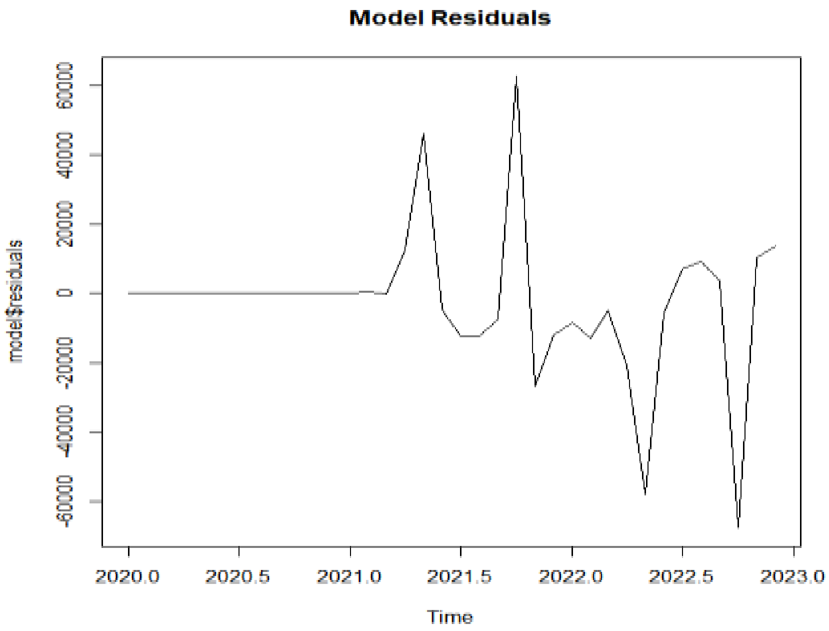


Figure 8: Model Residuals

4.8 Normality

Below is the histogram showing the normal curve obtained from the data after running the below commands using R- Studio. This shows that the data is normally distributed.

CODE 10 in R Studio

> hist (model$residuals, main = "Histogram Residuals”, border = "black”, probability = TRUE)

> lines (density (model$residuals), col = "red")

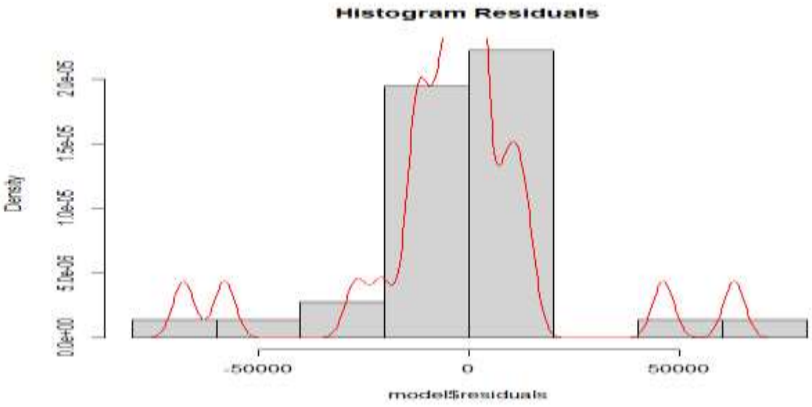


Figure 9: Histogram of Residuals

4.9 Normal Q-Q Plot Residuals

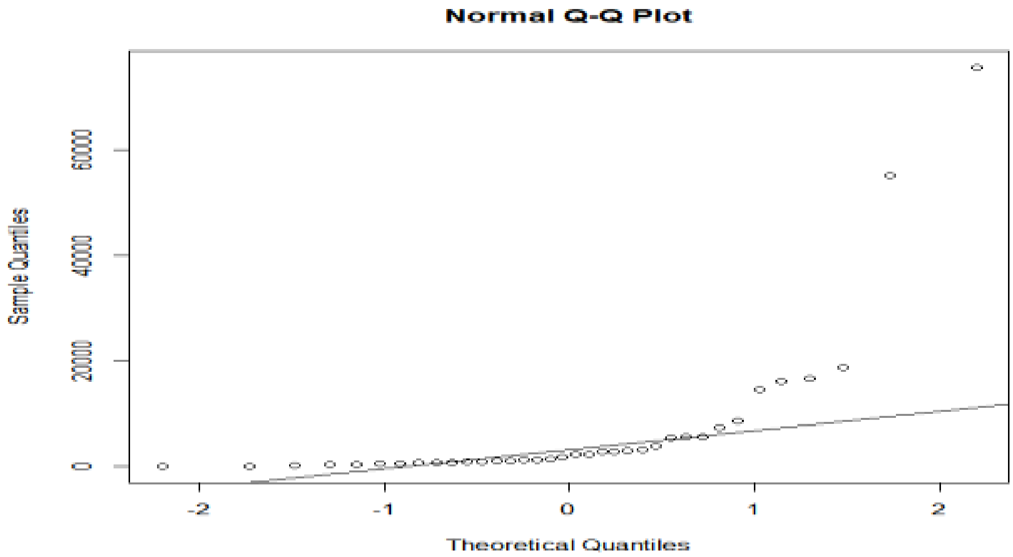
The Q-Q plot in Figure 10 provides a visual representation of the normality of the dependent variable. This is shown below since the points are generally in a
straight line.

CODE 11 in R Studio

> qqnorm (NUMBER.OF.COVID.19.IN.ZIMBABWE)

> qqline (NUMBER.OF.COVID.19.IN.ZIMBABWE)

Figure 10 Normal Q-Q Plot Residuals



4.10 Serial Autocorrelations

The researcher tested the following hypothesis for serial autocorrelation as below:

H_0: There is no serial autocorrelation.

H_1: There is serial autocorrelation.

4.10.1 Results of Box-L Jung Test

Data: model$residuals

CODE 12 in R Studio

> Box. test (model$resid, lag=3, type ="Ljung-Box")

Table 7: Box-L Jung test

Box-L Jung test
Data: model$resid
X-squared = 1.589
DF = 3
p-Value = 0.6619

Table 7 above shows the results of a Box-Ljung test performed on the residuals of a model, shown by ‘model$resid’. The X-squared value displayed is the test statistic, with a value of 1.589 in this case, and the degree of freedom is 3. The p-value represents the likelihood of obataining the test-statistic value, on the assumption that the residuals are autonomously distributed. In this case, the p-value of 0.6619 indicates limited evidence against the null hypothesis that the residuals are in dependently distributed, indicating that the model adequately accounts for the variability in the data. Since the p-value of 0.6619 is higher than the significance level of 0.05 for each of the tested lags, we accept the null hypothesis and conclude that there is no serial autocorrelation in the fitted model. The residuals are independent.

4.11 Time Series Data Forecasting For Ten Steps

The process of forecasting time series involves using past data or observations of the original time series to predict future values (Cryer and Chan, 2008).The previously observed COVID-19 confirmed values are used to forecast the future values for proper economic study and implementations. The researcher executed the following code to predict the period:

CODE 13 in R Studio

```
> Forecast = forecast (model, h = 8, level = c (95))
> plot (Forecast)
```

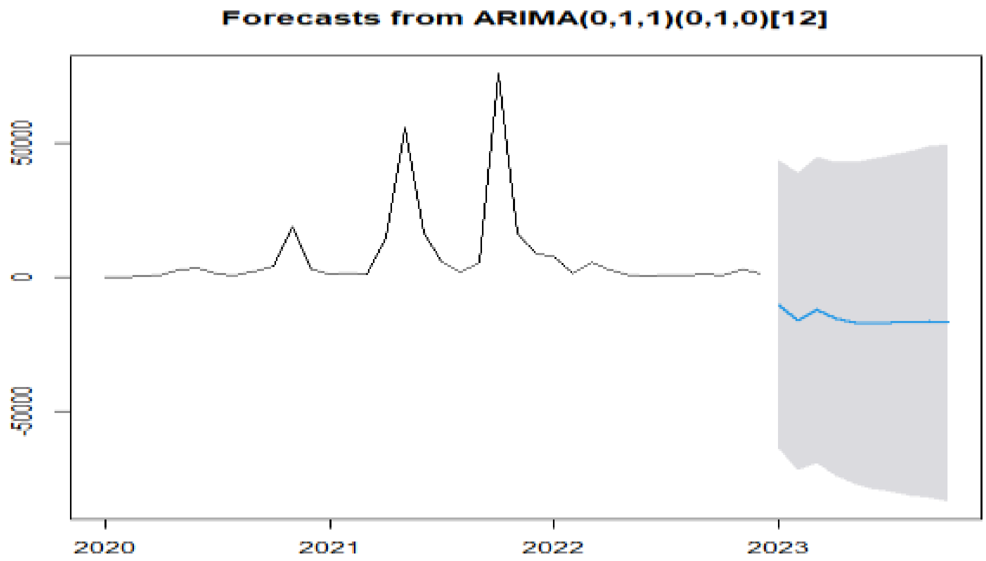


Figure 11: Observed Time Series and the Forecasted Future Values (2020 -2023)

Figure 11 shows the predicted future COVID-19 cases starting with a quick fluctuation, then a slight drop in 2019 followed by a constant rise to the same level as before. Afterwards, the future Number of COVID-19 confirmed cases decreases, and it will rise again up to October 2023 as indicated by the Figure 11. The predicted values for monthly forecasts were given below with 95% confidence Interval. The trend indicated that there is no increase in the Number of COVID-19 in Zimbabwe.

CODE 14 in R Studio

> Forecast

TABLE 8: Forecasted Number of COVID-19 cases

Point	Forecast	Lo 95	Hi 95
Mar 2023	-10305.78	-64096.38	43484.81
Apr 2023	-16406.78	71756.09	38942.52
May 2023	-12271.78	69137.10	44593.53
Jun 2023	-15395.78	73737.73	42946.16
Jul 2023	17099.78	76881.89	42682.32
Aug 2023	17390.78	78579.17	43797.60
Sep 2023	17005.78	79568.84	45557.27
Oct 2023	17115.78	81023.95	46792

The table shows the point forecasts, as well as the lower and upper bounds of the 95% prediction intervals, for the time series variable being forecasted. These values are given for each of the months from March 2023 to October 2023.

The point forecast for each month represents the predicted value of COVID-19 cases for that particular month. The Lo 95 and Hi 95 columns represent the lower and upper bounds of the 95% prediction interval for each point forecast for each point forecast, thereby indicating the degree of uncertainty in the forecast.

Thus, these values can be used as an estimate of future values and to determine a range of possible outcomes with a confidence level of 95%.

4.12 Summary

The analysis and presentation of data in this chapter allowed the researcher to identify the most appropriate time series model for the data. This was achieved through diagnostic checks performed on the models. This also helped to forecast Number of Covid-19 to be confirmed in Zimbabwe in the next 8 months of 2023 from March to October.

CHAPTER FIVE: SUMMARY, CONCLUSION AND DISCUSSION

5.0 Introduction

The chapter gives a summary of the time series analysis of Zimbabwe's COVID-19 confirmed cases and the recommendations drawn from the study.

5.1 Summary on the findings of the study

The researcher studied on the time series on Zimbabwe's COVID-19 confirmed cases from March 2020 up to February 2023. The researcher collected data and organized it into monthly intervals, while also utilizing data published by the World Health Organization (WHO). The purpose of the study was to make predictions regarding the number of confirmed COVID-19 cases in Zimbabwe. The literature review explains how COVID-19 affected worldwide including Zimbabwe and how it becomes a pandemic. Therefore, the researcher conducted a research on forecasting future Number of COVID-19 in the country in order to take measure which will make safety for citizens of Zimbabwe. In order to obtain best results, the researcher had to find the best fit time series model to forecast Number of COVID-19 confirmed cases in Zimbabwe in the next eight months. In the theoretical framework, the researcher outlined the ARIMA model and SARIMA models to predict the Number of COVID-19 confirmed cases in Zimbabwe water trends in the next eight months. The research engaged the descriptive research design. R-studio and Excel are the software used to run the data by the researcher.

The aim of forecasting the Zimbabwe's COVID-19 confirmed cases for the data was revealed in the Figure 11 and explained in Table 8. The researcher managed to come out with the best time series model to forecast Number of COVID-19 confirmed cases in Zimbabwe. The best time series model was ARIMA (0.1.1) (0.1.0) [12]. The researcher used several graphs to test for stationarity and also stationarizing the data to obtain a suitable model for the data. The researcher also test for normality using histogram this was shown in Figure 9. The researcher went on to forecast Number of COVID-19 confirmed cases in Zimbabwe using R-Studio and this is explained in the previous chapter. The forecasted data was for the period March 2023 up to October 2023. The graph showed a decreasing trend on the Number of COVID-19 confirmed cases from the period March 2023 to October 2023. This was because of natural immunity among the population due to prior exposure to the virus. The public awareness campaigns and initiatives to promote preventive measures. The implementation of strict public health measures, including lockdowns, social distancing, and mandatory mask-wearing, for a limited duration led to a decline in the number of confirmed COVID-19 cases in Zimbabwe. The final chapter of the research presents a summary of the research findings, draws conclusions based on those findings, and includes a discussion of the study.

5.2 Conclusion

From the findings of the research, we can conclude that the best fit time series model in Forecasting Zimbabwe's COVID-19 confirmed cases was ARIMA (0.1.1) (0.1.0) [12]. Since there was a decrease trend in forecasted data, it shows that for the next eight months the will be less or no cases will be confirmed in Zimbabwe.

5.3 Recommendations

On the basis of the findings of the research, the researcher recommends other scholars to conduct the forecasting process using different methods of forecasting to come up with the best fit time series model.

References

Australian Bureau of Statistics: Time Series Analysis: The Basics (https://www.abs.gov.au/websitedbs/d3310114.nsf/home/time+series+analysis:the+basics [accessed on 22 January 2023])

Cohort design.

Available online: (https://library.sacredheart.edu/c.php?g=29803&p=185902 [accessed on 5 February 2023])

COVID Live-Coronavirus Statistics-Worldometer.

Available online: (https://www.worldometers.info/coronavirus/[accessed on 9 February 2022])

Cryer, J.D., &Chan, K. (2008). Time series Analysis with Applications in R. lowa city USA: Springer Texts in Statistics.

Document and records.

Available online: (https://www.indeed.com/career-advice/career-development/reseach-methodology [accessed on 5 February 2023])

Da Hye, L.,Youn Su, K., & Hong, C. (2021). Forecasting COVID-19 Confirmed Cases using Empirical Data Analysis in Korea.

Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7998453/

Guo, G., Ye, L., Pan, K., Chen, Y., Xing, D., Yan, K. (2020). New insight of emerging SARS-CoV-2:epidemiology, etiology, clinical features, clinical treatment, and prevention. Front Cell Dev Biol. Front Cell Dev Biol. 2020 May 22; 8:410. PubMed / Google Scholar

Governmental of Zimbabwe. Zimbabwe COVID-19 operational plan May-July 2020.Harare, Zimbabwe: Ministry of Health and Child Care, 2020.

Gregory, C., & Gwilym, M. (2008). Time series analysis. Forecasting and Control. Canada: A John Wiley and sons, inc, publication.

Hove, B.,& Dube, B.(2021) COVID-19 and the entrenchment of a virtual Elite private school: Rethinking education policies in Zimbabwe. J. Cult. Values Educ. 2021, 4, 84-94.

Available online: https://www.cultureandvalues.org/index.php/JCV/article/view/126

Kavenga, F., Rick, H.M., Chinongo, R., Taruvinga, T., Marembo, T., Manasa, J., Marambire, E., McHugh, G., Greyson, C.L., Bandason, T. (2021). Comprehensive occupational health services for healthcare workers in Zimbabwe during the SARS-CoV-2 pandemic. PLoS ONE 2021,16, e0260261. [Google Scholar] [Cross Ref]

Remuzzi, A., & Remuzzi G. (2020). COVID-19 and Italy: What next ? Lancet. 2020; 395(10231): 1225-8. Google Scholar

Sakhabakhsh, L., & Yarmohammadi, M. (2012). An empirical study of the usefulness of SARFIMA models in science.

Available online: https://www.researchgate.net/publication/260571504_An_empirical_study_of_the_usefulness_of_SARFIMA_models_in_energy_science

Srivastava, N., Baxi, P., Ratho, R.K., & Saxena, S.K.(2020). Global trends in epidemiology of coronavirus disease 2019 (COVID-19).Coronavirus Disease 20219 (COVID-19).2020:9-21. PubMed / Google Scholar

Tsay, R.S (2010). Analysis of Financial Time Series: Third Edition. In Analysis of Financial Time Series: https//doi.org/10.1002/9780470644560.

Madziva, R., Murewanhema, G., Dzinamarira, T., Herera, H.,& Musuka, G. (2021).Enhancing SARS-CoV-2 surveillance at ports of entry between South Africa and Zimbabwe due anticipated increased human mobility during festive period. Public Health Pract.Oxf.Engl.2021,2,100215.[Google Scholar] [Cross Ref]

Makoni, T., & Chikobvu, D. (2018). Modelling and Forecasting Zimbabwe's Tourist Arrivals Using Time Series Method: A Case Study of Victoria Falls Rainforest.

Available online: https://doi.org/10.25159/1998-8125/3791

Mapuwei, T.W., et al (2020). Univariate Time Series Analysis of Short-Term Forecasting

Murewanhema, G., & Mutsigiri-Murewanhema, F.(2021). Drivers of the third wave of COVID-19 in Zimbabwe and challenges for control: Perspectives and recommendations. Pan Afr. Med. J. 2021, 40, 46.

Available online: https://www.panafrican-med-journal.com/content/article/40/46/full

Murewanhema, G., Burukai, T.V., Chiwaka, L., Maunganidze, F., Munodawafa, D., Pote, W., & Mufunda, J.(2021). The effect of increased mobility on SARS-CoV-2 transmission: A descriptive study of the trends of COVID-19 in Zimbabwe between December 2020 and January 2021. Pan Afr. Med. J. 2021, 39, 125.

Montgomery, D. C., Jennings, C., & Kulahci, M. (2015). Introduction to Time Series Analysis and Forecasting (Second Edition).

Partial Autocorrelation Function (PACF)

Available online: (https://online.stat.psu.edu/stat510/lesson/2/2.2 accessed on 2 February 2023)

Perone, G. (2022). Using the SARIMA model to forecast the fourth global wave of cumulative deaths from COVID-19

Providing a Learning Solution for Millions of in and out of School Children in Zimbabwe. Available online: (https://www.unicef.org/zimbabwe/stories/providing-learning-solution-millions-and-out-school-children-zimbabwe [accessed on 12 March 2022])

World Health Organization. Coronavirus disease(COVID-19) situation report 163. 1st July 2020.

World Health Organization. COVID-19 WHO African region: external situation report 18/2020 Available online: https://www.panafrican-med-journal.com/content/article/39/125/full [Cross Ref]

World Bank. Zimbabwe Economic Update: COVID-19 Further Complicates Zimbabwe’s Economic and Social conditions

Available online: (https://www.worldbank.org/en/country/zimbabwe/publication/zimbabwe-economic-update-covid-19-further-complicates-zimbabwe-s-economic-and-social-conditions [accessed on 9 February 2022]

Zinyemba, L., Nhango, K., & Zinyemba, A.(2021). COVID-19 induced online learning : The Zimbabwean experience. Afr. J. Soc. Work 2021, 11, 223-230. [Google Scholar]

Zhao, D. (2022). The research of ARIMA, GM(1,1) and LSTM models for prediction of TB cases in China.

Available online: https://pubmed.ncbi.nlm.nih.gov/35196309/

Appendix

Descriptive summary

Mean	7340.833
Standard Error	2556.418
Median	2055
Mode	#N/A
Standard Deviation	15338.51
Sample Variance	2.35E+08
Kurtosis	13.3555
Skewness	3.574992
Range	75683
Minimum	8
Maximum	75691
Sum	264270
Count	36

```
> adf.test(NUMBER.OF.COVID.19.IN.ZIMBABWE)

          Augmented Dickey-Fuller Test

Data  :  NUMBER.OF.COVID.19.IN.ZIMBABWE

Dickey-Fuller = -2.499, Lag order = 3, p-value = 0.379
alternative  hypothesis: stationary

> A = diff(log(NUMBER.OF.COVID.19.IN.ZIMBABWE))
> adf.test(A)

      Augmented Dickey-Fuller Test

Data  :  A
Dickey-Fuller = -6.3923, Lag order = 3, p-value = 0.01
alternative  hypothesis: stationary
> library(ggplot2)
> tsdata = ts(NUMBER.OF.COVID.19.IN.ZIMBABWE,frequency = 12)
> ddata = decompose(tsdata,"multiplicative")
> plot(ddata,title(main = "Decomposition of the Data into Time-Series Components"))
> par(mfrow = c(1,2))
> acf(NUMBER.OF.COVID.19.IN.ZIMBABWE,main = "ACF")
> pacf(NUMBER.OF.COVID.19.IN.ZIMBABWE,main = "PACF")
```

```
> library(ggplot2)
> COVID = na.locf(NUMBER.OF.COVID.19.IN.ZIMBABWE,fromLast = TRUE)
> COVID = ts(COVID,start = 2020,frequency = 12)
> COVID
```

```
> auto.arima(COVID,D = 1,ic = "aic",trace = TRUE)

 ARIMA(2,1,2)(1,1,1)[12]                    : Inf
 ARIMA(0,1,0)(0,1,0)[12]                    : 546.818
 ARIMA(1,1,0)(1,1,0)[12]                    : Inf
 ARIMA(0,1,1)(0,1,1)[12]                    : Inf
 ARIMA(0,1,0)(1,1,0)[12]                    : Inf
 ARIMA(0,1,0)(0,1,1)[12]                    : Inf
 ARIMA(0,1,0)(1,1,1)[12]                    : Inf
 ARIMA(1,1,0)(0,1,0)[12]                    : 544.8943
 ARIMA(1,1,0)(0,1,1)[12]                    : Inf
 ARIMA(1,1,0)(1,1,1)[12]                    : Inf
 ARIMA(2,1,0)(0,1,0)[12]                    : 544.0708
 ARIMA(2,1,0)(1,1,0)[12]                    : Inf
 ARIMA(2,1,0)(0,1,1)[12]                    : Inf
 ARIMA(2,1,0)(1,1,1)[12]                    : Inf
 ARIMA(3,1,0)(0,1,0)[12]                    : 544.4576
 ARIMA(2,1,1)(0,1,0)[12]                    : 542.5244
 ARIMA(2,1,1)(1,1,0)[12]                    : Inf
 ARIMA(2,1,1)(0,1,1)[12]                    : Inf
 ARIMA(2,1,1)(1,1,1)[12]                    : Inf
 ARIMA(1,1,1)(0,1,0)[12]                    : 541.1966
 ARIMA(1,1,1)(1,1,0)[12]                    : Inf
 ARIMA(1,1,1)(0,1,1)[12]                    : Inf
 ARIMA(1,1,1)(1,1,1)[12]                    : Inf
 ARIMA(0,1,1)(0,1,0)[12]                    : 539.2184
 ARIMA(0,1,1)(1,1,0)[12]                    : Inf
 ARIMA(0,1,1)(1,1,1)[12]                    : Inf
 ARIMA(0,1,2)(0,1,0)[12]                    : 541.1839
 ARIMA(1,1,2)(0,1,0)[12]                    : Inf
```

Best model: ARIMA(0,1,1)(0,1,0)[12]

```
> model = auto.arima(COVID,D=1)
> summary(model)

Series: COVID
ARIMA(0,1,1)(0,1,0)[12]

Coefficients:
          ma1
      -0.7575
s.e.   0.1301

sigma^2 = 753210228:  log likelihood = -267.61
AIC=539.22   AICc=539.82   BIC=541.49

Training set error measures:
                    ME      RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set  -2460.809  21454.49  11649.91  -491.5673  734.8522  0.7511732 -0.01788781
```

```
> plot.ts(model$residuals,main = "Model Residuals")
> hist(model$residuals,main = "Histogram Residuals",border = "black",probability = TRUE)
> lines(density(model$residuals),col = "red")
> qqnorm(NUMBER.OF.COVID.19.IN.ZIMBABWE)
> qqline(NUMBER.OF.COVID.19.IN.ZIMBABWE)
> Box.test(model$resid, lag=3, type ="Ljung-Box")
> Forecast = forecast (model, h = 8, level = c (95))
> plot (Forecast)
> Forecast
```