MODELLING DAILY COVID-19 CASES IN KENYA USING ARIMA AND SARIMA MODELS

CAROLINE MUTHONI KAMOTHO

PA200/S/12175/20

A Research Project Submitted to the School of Pure and Applied Sciences in Partial Fulfillment of the Requirements for the Award of Masters of Science in Statistics of Kirinyaga University

KIRINYAGA UNIVERSITY

September, 2023

DECLARATION AND RECOMMENDATION

This project is my original work and has not been presented elsewhere for a degree award.

Signed:.. Date: ....

Kamotho Caroline Muthoni

PA200/S/12175/20

RECOMMENDATION

This research project has been submitted for examination with our approval as University supervisors.

Signed:.. Date: ....

Dr. Josephine Njeri Ngure, PhD

Department of Pure and Applied Sciences

Kirinyaga University

Signed:.. Date: ....

Dr. Margaret Wambui Kinyua, PhD

Department of Mathematics, Statistics and Actuarial Science

Karatina University

DEDICATION

To my parents Peter and Judy and my sister, Valencia.

ACKNOWLEDGMENT

I express my sincere gratitude to my supervisors; Dr. Josephine Ngure, PhD and Dr. Margaret Kinyua, PhD for their dynamism, vision, sincerity, and invaluable guidance throughout this research. I'm incredibly grateful to my parents and sister for their prayers and sacrifices in educating and preparing me for my future. They have given me not only the time to devote to the completion of this great challenge, but also the advise to continue and complete it. My friends and my classmates, who assisted me in my research, have been very supportive and encouraging. I also thank the School of Pure and Applied Sciences faculty and staff, who were always willing to assist me and point me in the right direction. Lastly and most importantly, I honor and glorify God, the almighty, for all the guidance and strength throughout my research work for successful completion.

ABSTRACT

Severe Acute Respiratory Syndrome is the primary cause of the pandemic coronavirus disease. The first case was reported in Wuhan, China, on 30^{th} December, 2019 with the first case on 13^{th} March, 2020 in Kenya. This contagious disease has become a global issue because it has resulted in millions of deaths, economic disruption leading to loss of employment and economic instability. This study therefore aimed at modelling daily COVID-19 cases in Kenya, using an Autoregressive Integrated Moving Average (ARIMA) model and a Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The specific objectives were: to fit an Autoregressive Integrated Moving Average (ARIMA) model, to fit a SARIMA model, to validate the model and to determine the forecast of COVID-19 cases. The World Health Organization was used as the source of secondary data dating from 13^{th} March, 2020 to 30^{th} April, 2023. These data was analyzed using R software. The training data was found to be non-stationary using a test known as Augmented Dickey Fuller, and it was differenced seasonally to make it stationary. The methodology used to fit the models was Box-Jenkins which uses the least AIC and BIC as its fitting criteria. The data revealed weekly seasonality hence invalidating the ARIMA model. SARIMA model was fitted and model validation using test data was done. The model with the least forecast errors was selected. The SARIMA(1,0,1)(2,1,2)$_7$ was selected with the least AIC = 2082.5, MAE = 2.9867, RMSE = 4.5815. Using the model, a ninety days forecast into the future was generated based on daily COVID-19 data. These forecasts will greatly create awareness of the trend and seasonality of this disease and therefore can be very useful to the health care providers as well as the government for purpose of planning, policy formulation, evaluation and resource allocation. This study recommends a comparative study on Bayesian SARIMA and SARIMA model to be perfomed, consideration of the possible change in probabilistic structures of the data and fitting of the BATS and TBATS models to the data.

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

ABBREVIATIONS & ACRONYMS

ACF	:	Autocorrelation Function
ACVF	:	Autocovariance Function
ADF	:	Augmented Dickey-Fuller Test
ADF$_T$	:	Augmented Dickey-Fuller Test statistic
AIC	:	Akaike Information Criterion
AR	:	Autoregressive
ARIMA	:	Autoregressive Integrated Moving Average
ARMA	:	Autoregressive Moving Average
BIC	:	Bayesian Information Criterion
COVID-19	:	Coronavirus disease of 2019
ETS	:	Educational Testing Service
MA	:	Moving Average
MCMC	:	Markov Chain Monte Carlo
MLE	:	Maximum Likelihood Function
MSE	:	Mean Squared Error
NAR	:	Nonlinear Autoregressive
NNAR	:	Neural Network Autoregression Model
NRMSE	:	Normalized Root Mean Squared Error
PACF	:	Partial Autocorrelation Function
PCR	:	Polymerase Chain Reaction
RMSE	:	Root Mean Squared Error
SARIMA	:	Seasonal Autoregressive Integrated Moving Average
SARIMAX	:	Seasonal Autoregressive Integrated Moving Average Exogenous
SARS-CoV-2	:	Severe Acute Respiratory Syndrome Coronavirus
SDGs	:	Sustainable Development Goals
SEIR	:	Susceptible Exposed Infected Recovered
WHO	:	World Health Organization

CHAPTER ONE

INTRODUCTION

1.1 Overview

This chapter outlines the project background, the statement of the problem, the main and specific objectives of the study, scope and, significance of the study.

1.2 Background of the study

1.2.1 COVID-19

On 11^{th} February, 2020, the term COVID-19 was coined by WHO to refer to the coronavirus disease. It is a contagious disease caused by a Severe Acute Respiratory Syndrome Coronavirus 2 known as the SARS-CoV-2 virus. This contagious virus is known to come from the large family of coronavirus (CoVs).

The International Committee on Virus Taxonomy adopted the name SARS-CoV-2 after the genetically related SARS-CoV. WHO uses COVID-19 to refer to SARS-CoV-2 to avoid confusion with the disease SARS. Polymerase Chain Reaction (PCR) reverse transcriptase is the test used to check whether the virus is present in a host. Minimal to moderate symptoms, and full recovery with zero treatment is what most infected people experience. However, there are individuals who need immediate medical consultation because of severe symptoms. Demorgraphic differences and pre-existing conditions have been the biggest risk factors, with individuals of advanced age having a higher chance of developing this illness. Although, any other individual can be infected with COVID-19 despite their age. Despite the fact that it can take upto fourteen days for an individual to be virus infected, symptoms can emerge within five to six days (Lima *et al.*, 2020).

Recurrent moderate signs and symptoms include coughing and high fever, exhaustion, and loss or reduced taste/smell. Symptoms can be mild in the form of; sore throat, body pains and aches, headaches or diarrhoea, a skin rash, fingers or toes discoloration and red eyes or eye irritation (Struyf *et al.*, 2022). Patients with severe symptoms like breath shortness, speech or mobility loss and chest pain are advised to immediately consult a doctor. The constant mutation of SARS-CoV-2 is a fact that cannot be disputed. Since the start of the pandemic, a number of notable variants have emerged which are, Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2) and Omicron (B.1.1.529) (Duong, 2021).

The contagious nature of this disease makes it spreadable through breathing in or by mouth, nose or eye contact after touching an infected surface (Jayaweera *et al.*, 2020). This virus is transmitted more rapidly in open spaces and windy conditions (Coskun *et al.*, 2021).

1.2.2 COVID-19 Vaccines

Even though the process of creating and manufacturing vaccines is laborious and time-devouring, several vaccines have been developed against this virus. To prevent virus infection, vaccines act like conventions which are typically administered to groups of healthy individuals, and not to unhealthy individuals to aid in their recovery. The types of vaccines that have been used are AstraZeneca, Pfizer-bioNTech, Johnson and Johnson, Moderna and Sinopharm. The post-authorization protection profiles of the COVID-19 vaccines have not revealed any indications of unexpected negative or harmful effects (Al Khames Aga *et al.* 2021). Being well informed with proper prior infor-

mation about the virus, its modes of transmission, treatment and diagnostic measures is the best strategy to stay safe from it and slow down the infection rate.
Kenya received the first batch of the Atrazeneca vaccine on 3^{rd} March, 2021 and thereafter administration of the COVID-19 vaccine began with a view to mitigate the spread of the virus thereby protecting Kenyan population.

1.2.3 Time Series Data and Models

Data about confirmed cases was usually collected and recorded on a daily basis per country and the information was shared to the World Health Organization. This data was then made available for researchers to analyze and find its characteristics as secondary data (Vartanian, 2010). According to Anderson (2011), time series refers to a variable defined as values that follow a sequence which is ordered based on intervals of time which are equally spaced. This time series data which also known as time-stamped data according to Naqvi *et al.*, (2017) can either be stationary with no mean and variance systematic changes or non-stationary. For time-stamped data to be non-stationary, it can be cyclic, seasonal, or contain a trend. The components can appear together. Time series models are usually fitted to data recorded over a period of time. The following are some models that are usually fitted and used for analysis:

i **AR(p)** model is a union of previous observations together with an error term.

ii **MA(q)** model is a merger of present and past/previous values of the random error term.

iii **ARMA(p,q)** model is an amalgamation of two models, AR and MA. It uses previously identified values alongside errors as the basis for future predictions.

iv **ARIMA(p,d,q)** When using non-stationary data, differencing is done to make the data stationary which results to an ARIMA model. The number of times data is differenced is represented by d in the model.

SARIMA is the preferred model to be fitted when data exhibits seasonality. SARIMA shows a spike at lag s in the ACF, where s is the period of seasonality. Time-stamped models are commonly applied in business, economics and finance. Time-stamped data models are used to perform analysis and generate possible predictions into the future (Naqvi *et al.*, 2017). Box-Jenkins methodology is the commonly used approach to generate these time series models from available data (Dritsakis & Klazoglou, 2018). This methodology is best used for time indexed historical data collected over a vast period of time (Ho *et al.*, 2002). ARIMA model has been fitted by Samson *et al.*, (2020) and Swain *et al.*, (2020) among many researchers for different aims or objectives. After analysis of time-stamped data and fitting of the model, forecasting is done. This is simply predicting or estimating future occurrences of COVID-19 infections based on the accessible historical data. Researchers have fitted most of these models to the COVID-19 data and generated forecasts. In the field of economics, SARIMA has widely been applied in forecasting, however, COVID-19 infections in Kenya has not been widely looked into.

1.3 Statement of the Problem

Globally, the new coronavirus (COVID-19 pandemic) continues to be a serious issue that impacts all facets of human endeavours. It is one of the most dangerous diseases to world public health, posing an unsettling scenario with more than six million deaths. The measures that were put in place to control its spread include lockdown, travel bans, gathering bans, and social isolation. These measures have had a huge negative impact on people worldwide for about two years. As a result, millions of people fell into extreme poverty. One of the SDG's is good health and well being and the amount of time estimated to achieve it has been elongated due to the sudden hit of this pandemic. One of the Big 4 agenda like Affordable healthcare and most of the other SDG's like no poverty and zero hunger and were also indirectly affected by the effects of this pandemic for example loss of employment. In previous studies, most researchers have analyzed a short COVID-19 dataset then forecasted the disease cases using ARIMA models in Kenya. Use of a long dataset increases the possibility of exploration, accurate results and identification of a seasonal component. However, data over a long period of time has not been analyzed using both ARIMA and SARIMA models in Kenya. The SARIMA model is a useful methodology for analyzing and forecasting daily COVID-19 cases in Kenya because it allows for the incorporation of lengthy data with seasonality. The availability of this long dataset allowed for the analysis and forecasting of COVID-19 using ARIMA and SARIMA models. Due to this reason and the continued reported COVID-19 infections, there was therefore a need for the Kenyan government to execute preparedness and ensure that sufficient resources were made available to combat the incidence of COVID-19. This research aimed at developing ARIMA and SARIMA models using recorded cases from 14^{th} March, 2020 to 30^{th} April, 2023 in Kenya and then forecast for 90 days into the future.

1.4 Research Objectives

1.4.1 Main Objective

The aim of this research was to model the daily COVID-19 cases in Kenya from 14^{th} March, 2020 to 30^{th} April, 2023 using ARIMA and SARIMA models.

1.4.2 Specific Objectives

The specific objectives of this research were;

i. To fit an ARIMA model to the daily COVID-19 cases in Kenya.

ii. To fit a SARIMA model to the daily COVID-19 cases in Kenya.

iii. To validate the SARIMA model using forecast errors.

iv. To determine forecast cases for 90 days using the SARIMA model.

1.5 Research questions

The research questions of this study were,

i. What was the best ARIMA model to fit COVID-19 data?

ii. What was the best SARIMA model to fit COVID-19 data?

iii. How valid was the selected SARIMA model?

iv. What were the generated forecast cases using a SARIMA model?

1.6 Scope of the study

This study aimed at fitting an ARIMA model, then fitting a SARIMA model to daily cases of COVID-19 to perform forecasting. This was because of possible seasonality in the data. The data that was analysed comprised of the daily number of COVID-19 infections at the beginning of March 2020 to July 2023 in Kenya.

1.7 Significance of the study

The relevance of this research in forecasting new COVID-19 cases is demonstrated in its applicability to make predictions of cases for any future pandemics. This will then translate to creation of preparedness for future pandemics.

The Kenyan government and the stakeholders can use this research to examine the possible disease burden during the pandemic which can affect the country's capability to achieve SDG's and Affordable housing in the Big 4 agenda on the set time frame.

1.8 Limitations of the study

Unreported cases steming from negligence or inadequate testing, the ommission of inteventions measures such as vaccinations, were some of the limitations of this study. Due to this, the focus was on the officially reported COVID-19 infections. Also, possible changes in probabilistic structures of the data for the period before and after restrictions were lifted was not considered.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter highlights a comprehensive literature review of previous studies. This includes studies of fitting time-stamped data models such as ARIMA and SARIMA. Further, the research gap was identified.

2.2 Theoretical review

2.2.1 ARIMA model

Samson *et al.*, (2020) conducted a research using a COVID-19 forecasting model in Nigeria. Using the confirmed COVID-19 cases, the study built an ARIMA model using the Box-Jenkins as the prediction methodology. On the differenced log-transformed data, stationarity was tested using ADF test. Using R^2 and normalized BIC, a comparison was performed. A two weeks forecast was generated by ARIMA (2,1,0) whose results had shown better performance than the other proposed models. Lockdown relaxation would lead to an increase in COVID-19 cases as indicated by the forecasts. With a sustained increase of these cases, a collapse of the health system would be iminent. Therefore, the Nigerian government needed to retain COVID-19 preventive measures.

Swain *et al.*, (2020 showcased how deaths and confirmed cases of COVID-19 could be predicted using an ARIMA model following a surge of cases reported in Odisha and India. An uptrend was predicted by the model for the two weeks that followed, and this was consistent with the actual recorded cases within the two weeks. Using the findings, the government was able to formulate pandemic preparedness policies for the health care systems which would ensure better preparation of the medical professions in combatting pandemics.

Khan & Gupta (2020) in their forecasts of Indian cases of COVID-19 made use of NAR and ARIMA models using a 50 days forecasting period. The researchers used data from the Ministry of Health and Family Welfare (MoHFW) to first fit an ARIMA model followed by a NAR and then compared the accuracy of the predictions from the models. From the researchers results, the predicted cases showed an uptrend which was consistent with the actual cases. After achieving the highest R^2 and least BIC values, ARIMA (1,1,1) was found to be appropriate model.

Bhangu *et al.*, (2021) analysed confirmed monthly cases using machine learning algorithms, which helped to reveal seasonality and trend of the data containing COVID-19 cases. A two weeks forecast using ARIMA and SARIMAX predicted the trend in the spread of COVID-19 which supported health-care services. Suitability of the model was established using a low MAE. Despite the existence of non stationarity and uncertainty in the data, the results showed a rising trend in the number of cases unless disease containerization measures were undertaken. Although the researchers used ARIMA and SARIMAX models, the data used was over a short period of time.

Chakraborty *et al.*, (2022), used empirical review to evaluate several forecasting models which worked on the short term, to predict daily cases of COVID-19 for different nations. Through a practicle study focusing on predicting accuracy, evidence was provided showing how the pandemic cases cannot be accurately predicted. Despite this, the generated predictions were still useful for healthcare resources allocation. This research suggested that having less data and insufficiet evidence made forecasting and nowcasting a challenge.

Perone (2021) fitted (ARIMA) model, ETS, NNAR, Box–Cox transformation done through trigonometric exponential smoothing state space, TBATS. All of their feasible hybrid combination were analyzed. This was done with an aim to predict the number of hospitalized patients exhibiting mild symptoms and critical conditions during COVID-19's second wave in Italy. The website of the Italian Health Ministry, was used as the source of the analyzed data. Outcome of the analysis showed that hybrid combination models were more efficient in correcting the linear, non - linear, and seasonal disease outbreak patterns. Following the projections that COVID-19 hospitalized patients numbers would exponentially increase, the number of ICU beds required were twice as much and three times for the next ten and twenty days respectively.

These forecasts were proportional to the reported, indicating that hybrid models may enable the judgment of public health authorities, particularly in the short term.

Maurice *et al.*, (2021) carried out an ARIMA model research study to forecast the Kenyan COVID-19 infections collected from the Ministry of Health. The descriptive statistics and the scatter plot developed a linear relationship between COVID-19 cases. Using the lease AIC from the least of possible models, ARIMA (0,1,2) emerged as the best model for forecasting using R software. One hundered and fifty days forecast were generated using R and they revealed an uptrend in the in the number of infections but the curve flattened after the constant increase rate. The authors suggested that the Kenyan citizens should heed to the given guidelines and the government should promote more widespread testing, quarantine, and provide free face masks and other protective equipment for medical personel. This research suggested ARIMA model was the most superior model to predict for the COVID-19 cases. However, the data used was for a period of eight months hence a need to analyze a more lengthy data.

Dehesh *et al.*, (2020) fitted an ARIMA model from Jan 2020 to March 2020 COVID-19 cases in China, Thailand and South Korea. These data was secondary data and was obtained from John Hopkins University. Choosing the best model for prediction and forecasting was the main aim of this study. Analysis was done using R and Stata and the forecast results revealed that China and Thailand prevented COVID-19 from spreading. Dehesh *et al.*, (2020) suggested for other countries to study the measures that these countries took. The study used short data length and therefore could not identify seasonality. The research did not have any conclusive forecasts either.

2.2.2 SEIR MODEL

Wambua *et al.*, (2022) determined the indirect effects of COVID-19 pandemic using longitudinal analysis. This was perfomed by analyzing the immunization and outpatient care data in Kenya from Kenya's health information system. An accuracy analysis was performed to test the validity of the estimates that were made to account for data that was missing. In April 2020, the total number of outpatient visits experienced significant declines. The predictions revealed a recuperation of the health services to normal by 21^{st} March as the effects of COVID-19 took a turnaround. A conclusion was made that a dynamic and active treatment was required to reverse the effects. This was because of an indirect implication of health services by COVID-19.

Kiarie *et al.*, (2022) created an SEIR model that has four elements to simulate how individuals interact with each other in a dynamic manner under four different conditions: susceptible (S), exposed (E), infected (I), and recovered (R). However, forecasting was performed using ARIMA model. During the fourth wave, peak daily cases were observed to be the lowest. According to the sixty days prognostications, there was an upward trend in the COVID-19 cases. 26^{th} October, 2021 being the peak of the fourth wave, it had four hundred and fifty four(454) new infections and forty(40) people who required immediate attention and sixteen(16) ICU cases due to severity. The results of this analysis were essential for drafting the containment measures and strategies of the pandemic. It was also relevant for enhancing the healthcare workers readiness and preparedness alongside the policymakers.

Odhiambo *et al.*, (2020) wanted to check whether the relationship of the risk components was linear. Most statistical models were used to model and forecast but there were no mathematical models for modelling and prediction purposes. Odhiambo *et al.*, (2020) therefore focused on applying a generalized linear regression to achieve these gaps. This was perfomed to quench the government's desire to know about the speed of COVID-19 rate of transmission. Additionally, it would assist in effective reception and care. This study aimed at using the compound Poisson to perform modelling of the of Kenyan COVID-19 cases and forecasting. The process's model parameters were generated from the daily contacts and number flights with reported and confirmed infections.

Ultimately, this research advised the government of Kenya to distribute enough testing kits all over the country as they put in more effort to improve public awareness. The Kenyan government was also advised to increase the number of quarantine centers, hospital beds and well trained medical personnel with proper protective gear.

2.2.3 SARIMA model

Tan *et al.*, (2022) aimed at analyzing COVID-19 cases recorded on a daily basis in Malaysia from January 2020 to September 2021. This was because of the rise in COVID-19 cases in the country raising a need to forecast for curbing measures to be put in place. Analysed data was secondary daily number of cases obtained from the Ministry of Health. The analysis was carried out using SPSS but the test for stationarity was done using R language. Differencing was carried out twice to make the data stationary. A testing and training set were generated from the data set. The selection of the optimum SARIMA model was based on the least Root Mean Squared Error, Mean Absolute Error and Bayesian Information Criterion. Validation of the optimum model was performed using the Ljung-Box test. Results of the twenty eight (28) conducted forecasts revealed that cases of COVID-19 were going down. Based on the errors arising during forecasting, this research concluded that the SARIMA models can make accurate forecasts. The resercher recommended the study to be carried out with more data points.

Feroze (2020) analyzed COVID-19 data of four months for five countries using the ARIMA model and then forecasted for thirty(30) days. The analysis was done to see the effects of lockdown in the five countries in different continents. The resulting forecasts of this study revealed an increasing trend in the COVID-19 cases alongside the resultant deaths during the lockdown measures easing period. Feroze (2020) used a short duration dataset and did not explore the SARIMA model.

As South Africa was trying to halt Malaria cases, Ebhuoma *et al.*, (2018) identified a need to find a goodmodel to forecast for the Malaria cases. This would play a big role in the control of Malaria cases. This study aimed at fitting the best SARIMA model to Malaria cases and use it to forecast. Testing and training sets were generated from the data set. The Box-Jenkins methodology was used for training and selection of the model. Forecasting was done using the model selected which was later validated against actual data. The model's forecasts closely fitted the actual cases. Ebhuoma *et al.*, (2018) then concluded that the model was good to be used to forecast for the Malaria cases. This would in return help in coming up with measures to curb the disease.

Hu *et al.*, (2007 wanted to develop a forecasting model and a model that could examine the relationship between temperature and cryptosporidiosis transmission. Analyzed data was secondary and it ranged from 1^{st} January, 1996 to 31^{st} December, 2004 for comparison. Generated results from both models highlighted the relationship occurring between weather fluctuation (temperature to be specific) and the rate of transmission of cryptosporidiosis. The least RMSE and AIC values were used as the criteria for selecting the best SARIMA model. The residuals from the model selected were eqivalent to the residuals assumptions from the model of best fit model.

Valipour (2015) evaluated the performance of two models, SARIMA and ARIMA on the forecasting relative errors. The long-term runoff data used was secondary data from 1901 to 2011 from all states stations in the United States of America. Analysis was done in two stages using two partitions of the data, first one was average runoff for each state and the second one was average runoff for all over the country. AIC values were used as the model selection criteria with parameter estimation carried out using the MINITAB software. The generated indicated an increase in temperature and a decreasing trend in rainfall. The SARIMA model produced more accurate forecasts compared to the ARIMA model with a relative error, RE < 5%. Comparing the two models, the research concluded that the SARIMA model performs better in forecasting. Although a hybrid SARIMA has been recommended, a SARIMA model had not been fitted to the daily COVID-19 cases in Kenya.

Perone (2022) forecasted for mid term to short term cumulative deaths of COVID-19 in twelve(12) different countries. This research used secondary data from Our World in Data to fit ARIMA and SARIMA models. The data-set was divided into training and test data set. The two models both proved that they were more accurate in forecasting compared to regular simple forecasting models. The SARIMA model performed better than the ARIMA model revealing a seasonal pattern in the data. This was deduced from the MAE and RMSE, where SARIMA model had the least values. The SARIMA model was then validated using data from 21^{st} August to 19^{th} Sept 2020 and it was found to be valid. Perone (2022) therefore concluded that SARIMA model was good for forecasting. This has however not been done in Kenya using more data points for COVID-19 cases.

2.3 Research gap

Most researchers in the previous studies, fitted ARIMA and SARIMA models to few COVID-19 data points. There was few researches that had been done to forecast the cases of COVID-19 using the ARIMA and SARIMA models in kenya. Due to the short period of data consideration, it was not possible to identify seasonality. With availability of more data points, there was therefore a need to fit ARIMA and SARIMA models incorporating the period: after the restrictions were lifted. This was because, arguably, the longer a time series is, the more accurate are the results of analysis.

This research therefore aimed at filling this gaps by using ARIMA and SARIMA models on a wider time frame data that would allow testing for model accuracy, test seasonality and forecast for COVID-19 cases in Kenya.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter mentions the target population as well as the source of the data. Data analysis method is clearly outlined where the model to be fitted is given. The related measures are derived as well as the procedure for model checking and validation.

3.2 Fitting ARIMA and SARIMA Models

The target population is inclusive of all positive COVID-19 infections in Kenya. This secondary data is sourced from the WHO's website and is analysed using R-Studio software. As part of the analysis of the time-stamped data, the first and foremost step is to plot data just as it was, generating a timeseries plot which is important in description of data (Moskovitch & Shahar, 2015). By observation, the plot helps identify features such as trend, seasonality, outliers and discontinuity in the data. The time series plot is also accompanied by some descriptive statistical measures to ensure that the structure of the data used in this research is well understood.

Trend and seasonality components could be found in time series data. Trend (T_t) is a gradual upward or downward movements due to factors that affect the mean of the series (Bee Dagum & Bianconcini, 2016). Seasonal variation (S_t) is a periodic movement in a series with a regularity of less than one year. It is possible to use a multipicative or additive model to combine and explain the previously mentioned components (Coghlan, 2018). Additive model is used to explain components that do not depend on time, they are roughly constant over time. Additive model is explained by the model given by Equation 3.2.1.

$$X_t = T_t + S_t + R_t \tag{3.2.1}$$
$$X_t = Trend + Seasonal + Random$$

In instances where the change in seasonality is directly proportional to the change in time, the multiplicative model may be an appropriate choice (Bhangu *et al.*, 2021). A multiplicative model formular is given by Equation 3.2.2.

$$X_t = T_t * S_t * R_t \tag{3.2.2}$$
$$X_t = Trend * Seasonal * Random$$

The data can be transformed to allow description using an additive model.
The following are the time series models that were fitted to the data.

3.2.1 White Noise

It is a collection of random variables $\{e_t\}$ which are usually uncorrelated with zero mean, and a finite variance (Bhangu *et al.*, 2021). This is a purely random process with no memory.
The mean, variance and covariance of white noise are given as 0, σ^2 and 0 respectively. White noise is not predictable because it has no memory hence very important in determining whether the model well fits the data. To test for white noise, the researcher used the visual tests, autocorrelation and normality test's or checking the autocorrelation function (ACF) of the errors using the correlogram.

3.2.2 Moving Average (MA) Model

The Moving Average process, which is commonly referred to as the MA model, is a basic time-stamped data model which is finitely stationary and is mostly used to model univariate time series data. Linearly combining past and present values of the error term of a white noise. A process X_t is said to be a moving average process of order q denoted by MA(q) if

$$X_t = \zeta_0 e_t + \zeta_1 e_{t-1} + \zeta_2 e_{t-2} + ... + \zeta_q e_{t-q} \tag{3.2.3}$$

where $e_t \sim \mathcal{N}(0, \sigma^2)$

$t = 1, 2, 3, ..., n$

Equation 3.2.3 can be simply written as;

$$X_t = \sum_{j=0}^{q} \zeta_j e_{t-j} \tag{3.2.4}$$

where ζ_j are constants or parameters to be determined.

3.2.3 Autoregressive (AR) Model

An autoregressive model is where the current observation can be written as linear combination of its p past observations together with the white noise (error terms). It is useful for prediction and inferencing. A process $\{X_t\}$ is said to an auto-regressive process of order p denoted by AR(P) if

$$\begin{aligned} X_t &= \delta_1 X_{t-1} + \delta_2 X_{t-2} + ... + \delta_p X_{t-p} + e_t \\ &= \sum_{k=1}^{p} \delta_k X_{t-k} + e_t \end{aligned} \tag{3.2.5}$$

where δ_t are the parameters, p is the observation lag number and,

$$e_t \sim \mathcal{N}(0, \sigma^2)$$

An AR model is not always stationary, it depends on the value of δ.

3.2.4 Autoregressive Moving Average (ARMA) Model

The AR combined with the MA model produces the ARMA model. Both the past observations and unexpected errors are considered. It was majorly introduced because it reduces the number of parameters used and it is defined by ARMA(p, q) with p and q defined as the orders of the AR and MA models respectively (B. Choi, 2012). It can be written as

$$X_t = \delta_1 X_{t-1} + \delta_2 X_{t-2} + ... + \delta_p X_{t-p} + e_t + \zeta_1 e_{t-1} + \zeta_2 e_{t-2} + ... + \zeta_q e_{t-q} \quad (3.2.6)$$

$$X_t = \sum_{k=1}^{p} \delta_k X_{t-k} + \sum_{j=1}^{q} \zeta_j e_{t-j} + e_t \quad (3.2.7)$$

δ = AR's model parameter/coefficients,

ζ = MA's model parameter/coefficients,

p, q = order of AR and MA respectively,

e_t = error term or white noise, $e_t \sim \mathcal{N}(0, \sigma^2)$.

An ARMA is a stationary process hence the mean and variance are constants and it does not require differencing.

Rewriting Equation 3.2.6 as;

$$X_t - \delta_1 X_{t-1} - \delta_2 X_{t-2} - ... - \delta_p X_{t-p} = \zeta_1 e_{t-1} + \zeta_2 e_{t-2} + ... + \zeta_q e_{t-q}$$

One can use the backward shift operator to obtain Equation 3.2.6 $B^j X_t = X_{t-j}$

$$\left[1 - \delta_1 B^1 - \delta_2 B^2 - ... - \delta_p B^p\right] X_t = \left[1 + \zeta_1 B^1 + \zeta_2 B^2 + ... + \zeta_q B^q\right] e_t$$

$$\varphi(B)X_t = \vartheta(B)e_t \tag{3.2.8}$$

where, $\varphi(B)$ =$(1 - \sum_{k=1}^{p} \delta_k B^k)$

and $\vartheta(B)$ =$(1 - \sum_{j=1}^{p} \zeta_j B^j)$ Equation 3.2.8 can be simplified to

$$X_t = \frac{\vartheta(B)e_t}{\varphi(B)} \tag{3.2.9}$$

Stationarity of the ARMA model is dependent on the AR model parameters i.e it is stationary if $\varphi(B)$ is stationary.

$$E(X_t) = \frac{\vartheta(B)}{\varphi(B)}E(e_t) \tag{3.2.10}$$

since

$$e_t \sim \mathcal{N}(0, \sigma^2)$$

then

$$E(X_t) = 0 \tag{3.2.11}$$

If ARMA process X_t is weakly stationary, then its representation as an infinite moving average is possible. i.e $MA(\infty)$.

If ARMA process X_t is invertible, then it can be represented as a infinite autoregressive model $AR(\infty)$. Both the autocorrelation and partial autocorrelations of an ARMA(p, q) tails of or approaches zero as lag h increases. This makes them not informative when choosing the order of an ARMA(p, q).

According to Gebretensae & Asmelash (2021), the extended autocorrelation can be used for the order selection.

3.2.5 Stationarity

The lack of periodic variations, mean and variance systematic changes are the characteristics that determine stationarity of time-stamped data (Dickey, 2015). On the other hand, non stationarity is characterized by the existence of seasonal, trend, cyclic components or the combination of these components. The types of stationarity are strict and weak stationarity. Weak stationarity, also known as 2^{nd} order stationarity occurs when the mean and variance don't vary with time i.e $E(X_t)$ and $var(X_t)$ are constants (Brockwell & Davis, 2009).

The covariance function $cov(X_t, X_{t+h})$ is independent of time but dependent on lag h. To check for the stationarity of the time-stamped data, the autocorrelation function(ACF) can be used. Time-stamped data makes a good prediction, if it is stationary.

To test for stationarity:One way to check for stationarity is by observing the time plot and the correlogram (Dickey, 2015). Time plots show horizontal upward trend with the variance being a constant. Then check how time t and time $t + h$ are correlated. If there is correlation then there will be dependence between the observations. To have stationarity, differencing can be performed on non-stationary time-stamped data. The following methods were used to check for stationarity;

Unit root test: This is a test or a check for data stationarity (Pesaran, 2007). The existence of a unit root for the time-stamped data is the precondition of the null hypothesis of non-stationarity. The alternative hypothesis is that of the time-stamped data being stationary. The Dickey-Fuller and Augmented Dickey-Fuller test are examples of the unit root test. Dickey-Fuller Test and Augmented Dickey-Fuller Test (ADF): The Dickey-Fuller test checks for the exsistence of a unit root in data to determine its stationarity. Time series may be more complex because the error term might not be white noise, hence the development of the Augmented Dickey-Fuller (Mushtaq, 2011). A series of differencing values are added to the DF test to evolve it to ADF.

The ADF is simply a Dickey-Fuller test augmented with lags of dependent variables. It

is a statistical significance test based on hypothesis testing. A test statistic is computed resulting to a $p - value$ whose inference reveals whether there is stationarity or not.
Hypothesis:

H_0 : Data is not stationary

H_1 : Data is stationary

Rejection of the null hypothesis is ensured by this test as it assumes non-stationarity of the data. The decision criteria involves comparing $p - value$ and stated value of significance. The null hypothesis is rejected and a conclusion reached that the process is stationary if the stated level of significance is greater than the resulting $p - value$.

3.2.6 Differencing

This is the conversion process non-stationary time series data to be stationary. It is done by subtracting one value from another successive value for the minimum number of times until stationarity is achieved. One can difference as many times as possible as long as stationarity is yet to be achieved. According to Moh'dMussa & Saxena (2018), when differencing is used to account for trend it is known as regular differencing and when it is used to account for seasonality it is known as seasonal differencing.

Let X_t denote a time series, differencing is carried out as follows, The 1^{st} differences are;

$$\bigtriangledown^{(1)} X_t = X_t - X_{t-1}$$

where $t = 1, 2, 3, ..., n$

The difference computed using backward shift operator B

$$\bigtriangledown^{(1)} X_t = (1 - B)Xt \tag{3.2.12}$$

$$\text{where, } B^j X_t = X_{t-j}$$

The second differences may be computed from the 1^{st} differences

$$\bigtriangledown^{(2)} X_t = \bigtriangledown^{(1)} X_t - \bigtriangledown^{(1)} X_{t-1} \tag{3.2.13}$$

The general differencing expresssion is as follows;

$$\bigtriangledown^{m} X_t = \bigtriangledown^{(m-1)} X_t - \bigtriangledown^{(m-1)} X_{t-1} \tag{3.2.14}$$

where, $\bigtriangledown$ denotes the difference

$m = 1, 2, 3, ...$ is the order of the difference (Moh'dMussa & Saxena, 2018).

3.2.7 Autoregressive Integrated Moving Average (ARIMA) Model

Removing parameters causing non-stationarity involves performing differencing (Zhang, 2018). An ARMA model when subjected to differencing becomes an ARIMA (p, d, q) model. (p, d, q) is the order of the AR, differencing and MA model. (Prabhakaran, 2019) ARIMA(p, d, q) model is used on the non-seasonal data to predict future values based on past observations only.

The model is

$$X_t = c + \delta_1 X_{t-1} + \delta_2 X_{t-2} + ... + \delta_p X_{t-p} + e_t + \zeta_1 e_{t-1} + \zeta_2 e_{t-2} + ... + \zeta_q e_{t-q} \tag{3.2.15}$$

The basic time series model is the ARIMA(1,1,1).

The general ARIMA model is of the form, (Zhang, 2018).

$$W_t = c + \delta_1 W_{t-1} + \delta_2 W_{t-2} + ... + \delta_p W_{t-p} + e_t + \zeta_1 e_{t-1} + \zeta_2 e_{t-2} ... + \zeta_q e_{t-q} \tag{3.2.16}$$

where $W_t = \bigtriangledown^{(d)} X_t$ is the difference

$B^j X_t = X_{t-j}$

$$\varphi(B)W_t = \vartheta(B)e_t \tag{3.2.17}$$

$$\varphi(B) \bigtriangledown^{(d)} X_t = \vartheta(B)e_t \tag{3.2.18}$$

but

$$\bigtriangledown^{(d)} X_t = (1-B)^d X_t \tag{3.2.19}$$

Therefore

$$\varphi(B)(1-B)^d X_t = \vartheta(B)e_t \tag{3.2.20}$$

Roots in the unit circle which are in the AR operator makes ARIMA a non-stationary process of order (p, d, q).

3.2.8 Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

Since ARIMA does not support data with seasonality, it is not fit for analyzing time-stamped data that exhibits seasonality. This drawback makes the SARIMA model appropriate as it incorporates seasonality. A seasonal pattern is a periodic movement with a regular pattern of less than one year or within a year. The model is denoted by SARIMA $(p, d, q)(P, D, Q)_s$. Using the backward shift operator, it is written as follows

$$(1-\varphi_p B)(1-\mathit{\Phi}_P B^s)(1-B)^d(1-B^s)^D X_t = (1+\vartheta_q B)(1+\mathit{\Theta}_Q B^s)e_t \tag{3.2.21}$$

or

$$\varphi(B^s)\mathit{\Phi}(B)(1-B^s)^D(1-B)^d X_t = \vartheta(B)\mathit{\Theta}(B^s)e_t \tag{3.2.22}$$

where,

φ and ϑ are the model parameters

B is the backward shift operator

p is the order of AR terms

d is the number of differencing

q is the order of MA terms

P is the order of seasonal AR terms

D is the number of seasonal differencing

Q is the order of seasonal MA terms

s = length of the season

Making $[X_t]$ the subject of equation 3.2.22 one obtains,

$$X_t = \frac{\vartheta(B)\varTheta(B^s)e_t}{\varPhi(B^s)\varphi(B)(1-B^s)^D(1-B)^d} \tag{3.2.23}$$

Example of ARIMA$(1,1,1)(1,1,1)_4$ is yearly data with quarterly patterns, where $s = 4$ is the order of seasonality.

$$= (1-\delta_1 B)(1-\delta_1 B^4)(1-B)(1-B^4)X_t = (1+\zeta_1)(1+\zeta B^4)e_t \tag{3.2.24}$$

The SARIMA model's order can be obtained using the ACF and PACF plots according to (Koyuncu *et al.*, 2021). For example, For yearly data with an annual pattern, $s = 12$

i ARIMA$(0,0,0)(0,0,1)_{12}$ ACF-One peak at lag 12 with all others with the others not significant PACF-Slow decay in the seasonal lags(12,24,36,...)

ii ARIMA$(0,0,0)(1,0,0)_{12}$ ACF-Slow decay on the seasonal lags PACF-One significant peak at lag 12

The Box-Jenkins methodology was employed to fit the time series models.

3.3 Box-Jenkins Methodology

This method was used for model identification, estimation and prediction (Devi *et al.*, 2013). The procedure is as follows;

3.3.1 Model Identification

The first step was Data Preparation: A time series plot was plotted and data transformation using offset logarithm was performed to ensure stability of the variance. Offset log transformation occurs when a constant is added to the data points before applying the log transformation (William & Wei, 2006). Usually perfomed on data that contains many zero values. Test for stationarity was done and differencing was carried out accordingly.
The second step is Model Selection: During this stage, plots of the ACF and PACF were used for the identication of appropriate model's orders. Stated here were some scenarios of the ACF plot that were applied in the identification of ARIMA and SARIMA model's parameters (Dritsakis & Klazoglou, 2018).

i Very slow or no decay in the ACF : This was a suggestion that the data was non-stationary and has long-range dependency.

ii Exponential slow decay to zero in the ACF : This suggested an AR model and therefore, the PACF plot was used to identify the order of p.

iii When there was one or more spikes but the other spikes were not significant in the ACF: Suggested the MA model of order q, where q was the spike that the ACF plot cut off from.

iv Slow decay that started after a few lags in the ACF and PACF : Implied the possibility of an ARMA model.

v No significant spikes at all in the ACF: Suggested a White noise or no autocorrelation (Majorly used when testing the residuals).

vi Spikes at multiples of one lag in ACF and PACF plot: This implied the SARIMA model hence the orders P and Q are taken based on the spikes at fixed intervals.

The AIC and BIC were used for the selection of the best model. This depended on the quality of the model which was estimated by selecting the least values of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Morley *et al.*, 2018).

AIC is efficient with a minimax property and it assumes normality (Vrieze, 2012). It is a means of comparative model quality estimation from a given a set of models. Parsimony and goodness of fit is balanced by AIC while ensuring that the selected model is generalizable (Cavanaugh & Neath, 2019). Let L be the maximum likelihood estimator and k be the number parameter estimates of the model, then the AIC of the model is calculated as follows,

$$AIC = -2logL + 2k \tag{3.3.1}$$

Selection of the best model was done by picking the least value of AIC.

BIC was introduced by Schwarz (1978) for independent and identically distributed observations and linear models whose likelihood was assumed to be from the exponential family.

It seeks to find the perfect model by strictly penalizing models with many parameters. It can be written as

$$BIC = -2logL + k\ln(n) \tag{3.3.2}$$

where L is the likelihood function

k =total parameters

n =total observations

BIC is majorly used in finite models where the best model selection follows the smallest BIC value (Neath & Cavanaugh, 2012). However, BIC is considered as less efficient for large data sets (I. So & AM, 2009).

3.3.2 Estimation and testing

After choosing an appropriate SARIMA model for the daily COVID-19 data, estimation of parameter values was done using the Maximum Likelihood Estimator method (Perone, 2021). This was because it was the most suitable and feasible method as there are error terms which can be classified as random components emerging from measurement errors. Therefore the likelihood could be easily obtained. Considering the daily COVID-19 cases, for example, $x_1, x_2, x_3, ...x_n$ are from a density function $f(x, \vartheta)$ where ϑ was the unknown parameter. The likelihood function was given by;

$$L(\vartheta) = \prod_{i=1}^{n} f(x_i, \vartheta) \tag{3.3.3}$$

To determine the MLE of ϑ, the likelihood function was differentiated with respect to ϑ and equated to zero,

i.e

$$\frac{dL(\vartheta)}{d(\vartheta)} = 0 \tag{3.3.4}$$

For many parameters, say up to k, then the likelihood function contains k parameters

i.e

$$L(\vartheta_1, \vartheta_2, \vartheta_3, ..., \vartheta_k) = \prod_{i=1}^{n} f(x_i, \vartheta_1, \vartheta_2, \vartheta_3, ..., \vartheta_k) \tag{3.3.5}$$

To determine the MLE of $\vartheta_1, \vartheta_2, \vartheta_3, ..., \vartheta_k$, differentiate the likelihood function with respect to $\vartheta_1, \vartheta_2, ..., \vartheta_k$ respectively while equating to zero for maximization.

3.3.3 Diagnostic checking

In this step, there was checking and testing whether the model was adequate and valid using residual diagnostics (Pagan & Hall, 1983). A histogram and the ACF of the residuals were plotted and a test for autocorrelation was performed using the Ljung-Box test. According to Serra & Rodríguez (2012), to determine the goodness of the model, autocorrelation of its residuals was tested using the Box-Ljung test. In this test, if the *p-value* was greater than the stated level of significance, the no autocorrelation null hypothesis was not rejected (Marquez *et al.*, 2015).

The best picked model should have normally distributed and white noise residuals.

3.3.4 Forecasting

If the best model was found in step three, the model then proceeded to step four and was used for forecasting. In the case of a model inadequacy in step three(3) for some reasons, the researcher goes back and repeats the process until a satisfactory model was found.

3.4 Model validation

At this point, the splitting technique was used to validate the selected models against the test data set. The models were used in forecasting for the test data set and a comparison of the forecasts and the actual data of the test set was carried out. This is known as forecasting out of the sample. The accuracy of the forecasts was obtained from the difference between actual cases and their predictions (E. C. So, 2013). The method of Root Mean Squared Error and Mean Absolute Error were used to check the accuracy of COVID-19 forecasted number of cases because according to Chai & Draxler (2014), they are measured in the same units as the variables.

MAE and RMSE are calculated using the following formulae,

$$\text{MAE} = \frac{\sum_{i=0}^{p} |X_i - \hat{X}_i|}{p} \tag{3.4.1}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{p} (X_i - \hat{X}_i)^2}{p}} \tag{3.4.2}$$

where,

X_i= the true case value

$\hat{X}_i$= the case prediction

p= the total observed COVID-19 cases

The normalized RMSE was also used to confirm whether the RMSE value was good enough (Shcherbakov *et al.*, 2013).

$$\text{Normalized RMSE} = \frac{\text{RMSE}}{\text{Maximum actual value-Minimum actual value}} \tag{3.4.3}$$

If the normalized RMSE which ranges between 0-1 is closer to 0, the model is a good fit. The model with the least MAE and RMSE was chosen as the best model and used for forecasting for ninenty days.

3.5 Forecasting

The best selected model was used in forecasting the future COVID-19 cases.

ARIMA model's forecasting equation used was given by Equation 3.5.1

$$W_t = c + \delta_1 W_{t-1} + \delta_2 W_{t-2} + ... + \delta_p W_{t-p} + e_t + \zeta_1 e_{t-1} + \zeta_2 e_{t-2} ... + \zeta_q e_{t-q} \tag{3.5.1}$$

A prediction about W_t where

$W_t = X_t - X_{t-1}$ is the differenced version of original COVID-19 data

SARIMA model's forecasting equation used was given by Equation 3.5.2

$$X_t = \frac{\vartheta(B)\mathit{\Theta}(B^s)e_t}{\mathit{\Phi}(B^s)\varphi(B)(1-B^s)^D(1-B)^d} \tag{3.5.2}$$

where, X_t is the forecast value based on past observations.

$B^j X_t = X_{t-j}$ where B is the backward shift operator.

ϑ and $\mathit{\Theta}$ are the model parameters as defined by Equation 3.2.22.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This chapter gives a description and interpretation of the results of COVID-19 data recorded from 14^{th} March, 2020 to 30^{th} April, 2023. There were 1143 data points. The data was obtained from the World Health's Organizations website. R stastistical software was used for analysis. Graphs and summary statistics were used to illustrate and interpret the results of analysis.

4.2 Fitting the ARIMA model

Estimation of the ARIMA model yielded the results discussed in this section. Only 80% of the data was used to fit the models as. This percentage was chosen according to Vrigazova (2021) who suggested splitting data into 20/80. The 20% of the data was used in model validation which is a very important step when fitting time series models (LeBaron & Weigend, 1998). According to Baglaeva *et al.*, (2020), models fitted using splitting generate more accurate results.

4.2.1 Descriptive Statistics

These included a time plot and a histogram of the original time-stamped data decomposed data and descriptive statistics.

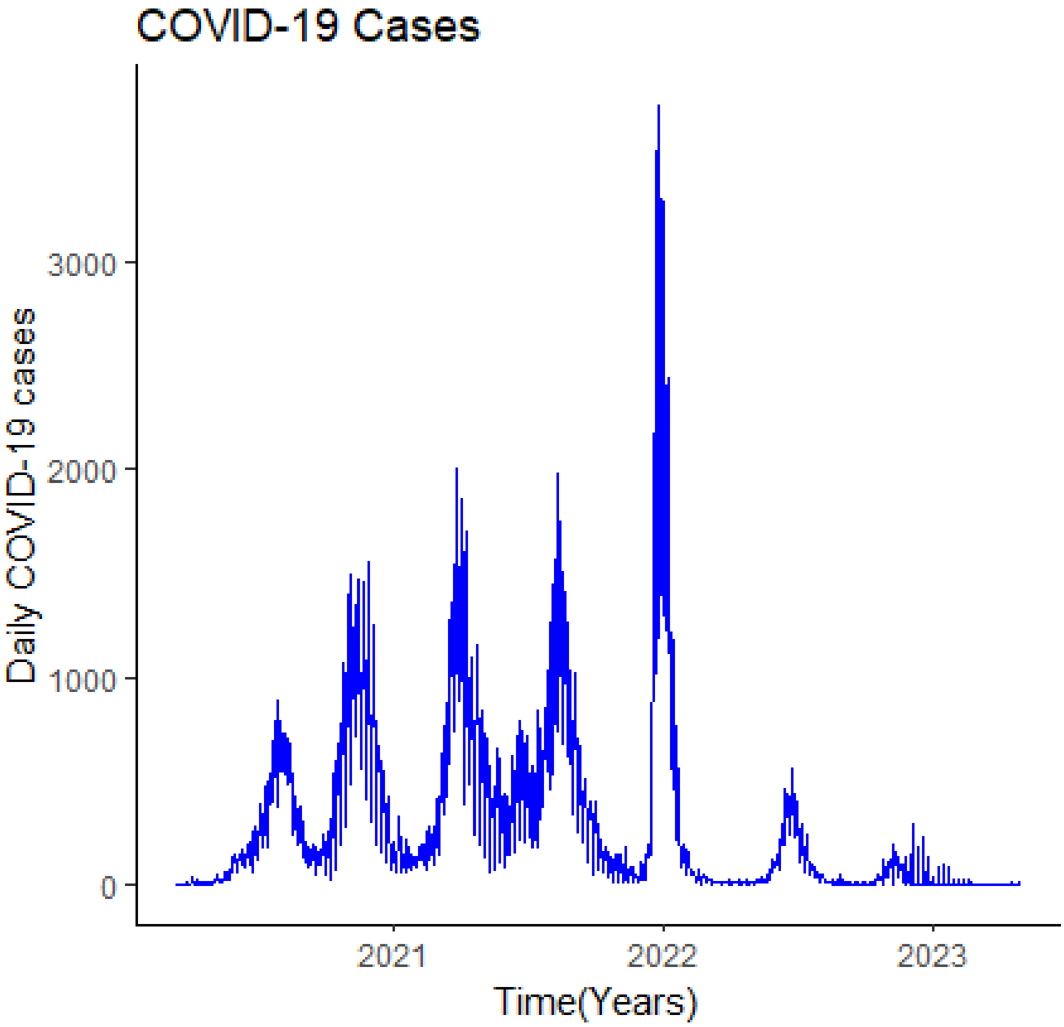

Figure 4.1: TimeSeries Plot of COVID-19 cases

Figure 4.1 revealed a high rise of daily COVID-19 cases between December 2021 and January 2022. This could have been caused by the lifting of restrictions by the Government of Kenya on 21st October, 2021. The freedom to travel and interact during the December festive season after a long period of lockdowns.

A rise in the number of cases led to the different peaks that indicated a difference in variances in the data. There were peaks at around July and November 2020, March, August and December 2021, and in June 2022. The highest peak was in December 2021 from which the trend decreased drastically with the maximum of 3749 COVID-19 cases recorded on 21st December, 2021. There were six troughs at around September 2020, February 2021, May 2021, October 2021, March 2022 and August, 2022. Troughs could have been due to the were due to the measures that were put in place to curb the spread and reduced travelling.

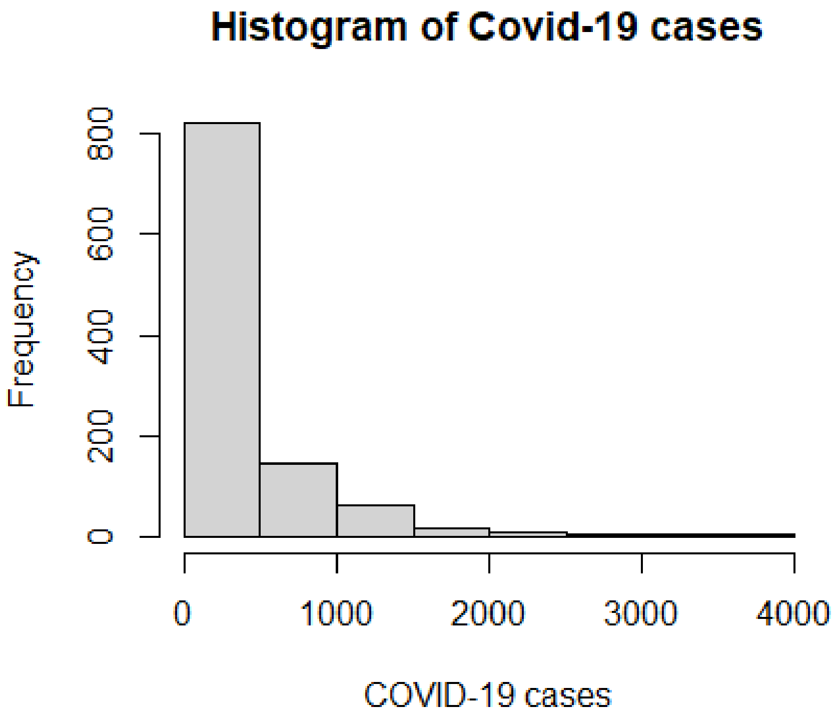

Figure 4.2: Histogram of COVID-19 cases

In Figure 4.2, the values with the highest frequency were $(0 - 1000)$. These were the cases that had been reported more often. The histogram was right skewed implying low frequency for cases between 1500 and 4000. Further, the non-normality of the histogram indicated different variances in the data, which would imply nonstationarity in the data set.

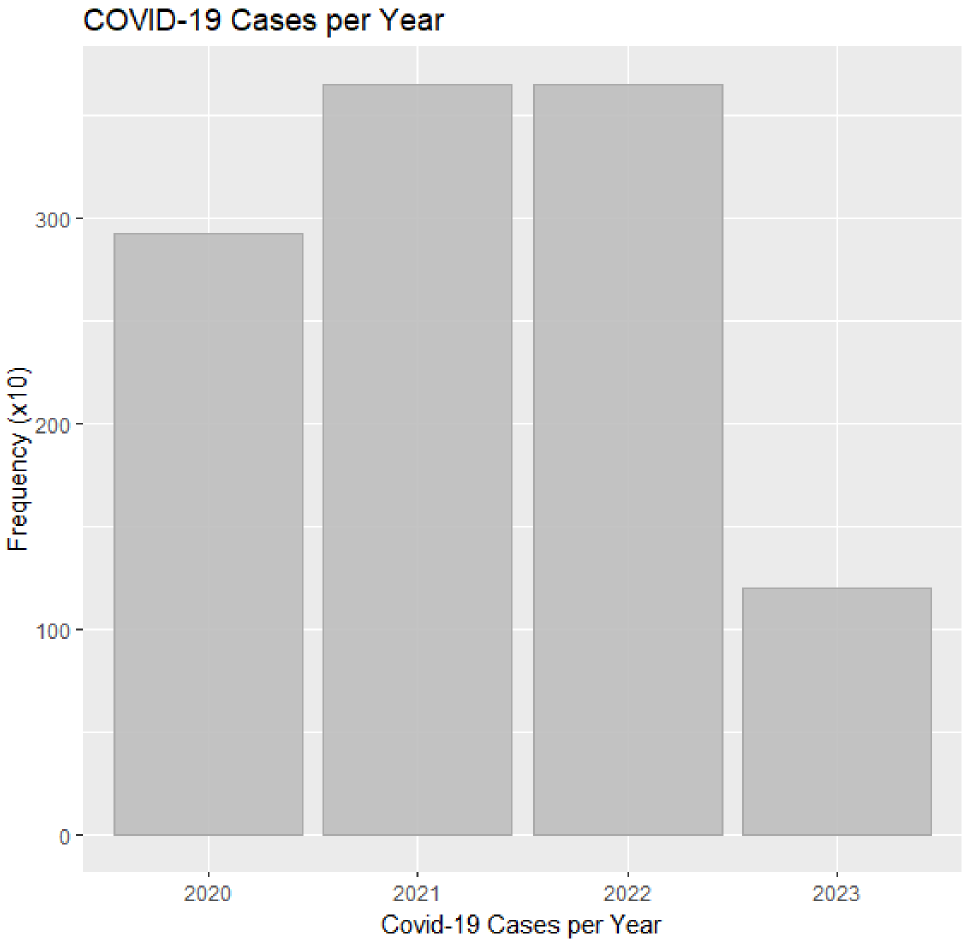

Figure 4.3: Histogram of COVID-19 cases per year

Figure 4.3 is slightly skewe to the right. The years 2021 and 2022 recorded the most number of COVID-19 cases. The disease then appeared to have been contained in the year 2023. This could have been due to the high numbers of vaccinations or the cases were no longer being reported. The cases had an increasing trend up until 2022 from which the cases started decreasing.

The COVID-19 data was decomposed and the individual time series components identified. These included trend, seasonal and random components.

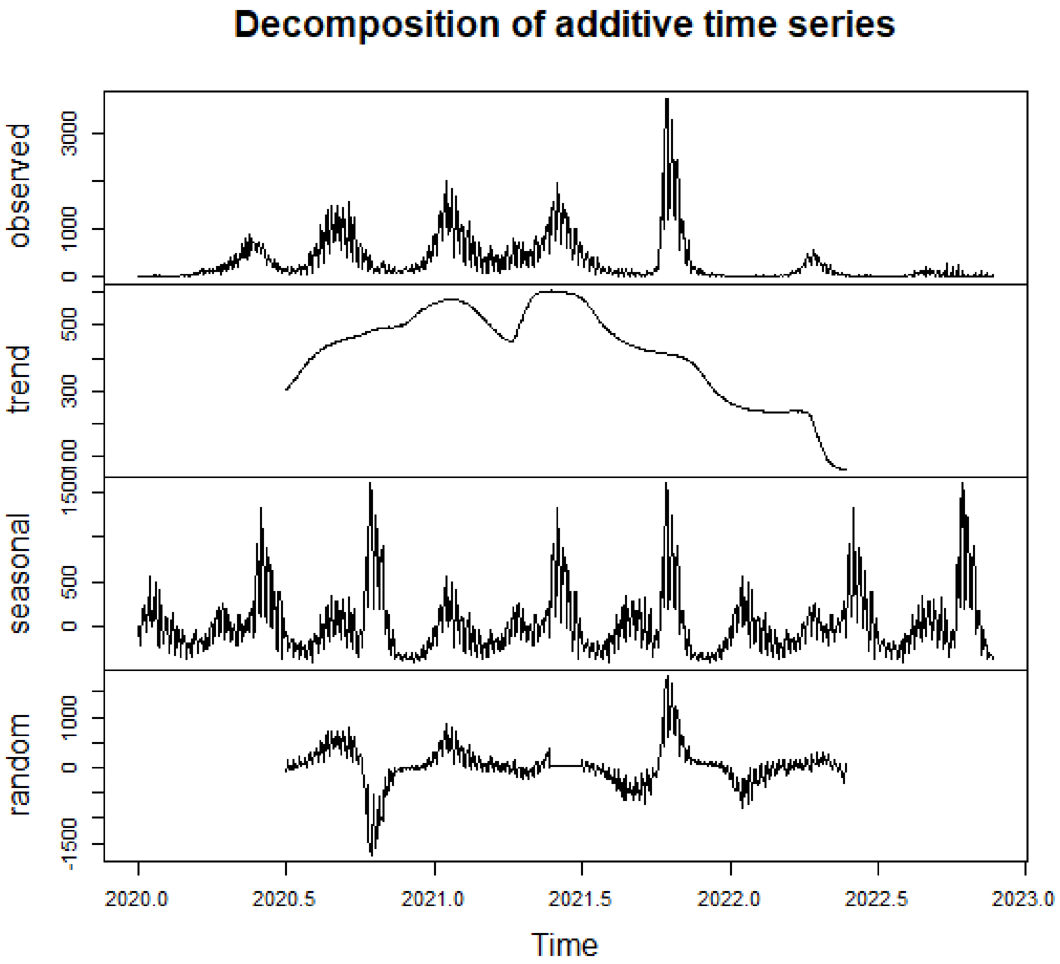

Figure 4.4: Decomposed plot of COVID-19 cases

Figure 4.4 showed the individual components of the COVID-19 data. There was an upward followed by a downward trend in COVID-19 infections. The data exhibited a seasonal component as seen in Figure 4.4. The data had a random component because of measurement errors such as false positives and inconsistencies in testing and reporting.

Table 4.1: COVID-19 cases descriptive statistics in Kenya for the period 2020-2023

Mean	Median	Skewness	Kurtosis	Shapiro-Wilk Test	Box-Ljung test
325	136	2.9393	12.3622	<2.2 $\times 10^{-16}$	<2.2 $\times 10^{-16}$

The average number of COVID-19 cases recorded per day was 325. Half the observations fell below 136 while half the observations fell above. In Figure 4.2, there was a positive skewness of 2.9393 and leptokurtosis of 12.3622. The Shapiro Wilk test produced $p - value < 2.2 \exp^{-16}$, an indication that normal distribution was not followed by the data (Razali *et al*., 2011). Autocorrelation(dependency) was tested using Box-Ljung test producing a $p-value < .05$ which indicated that the data was strongly correlated.

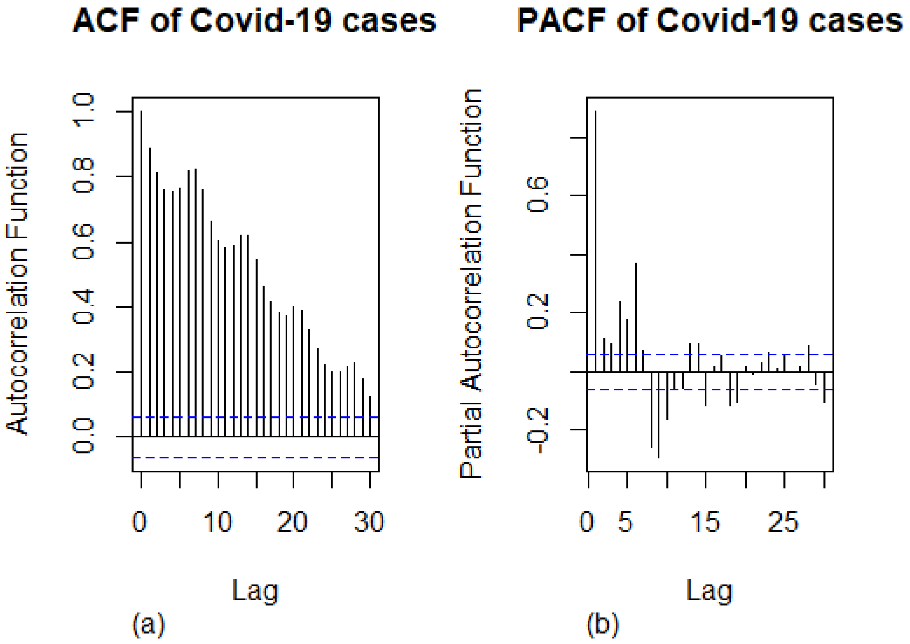

Figure 4.5: ACF and PACF of COVID-19 cases

Slow decaying pattern of the ACF plot in Figure 4.5(a) showed that the time series was not stationary. The slow decay was also an indication of long range dependency in the recorded COVID-19 cases. The data had significant positive correlations and therefore the time series data was not random since there were many non-zero spikes in the ACF plot. From the PACF plot in Figure 4.5(b), the data had a high degree of autocorrelation.

Since the data revealed that it had different variances and skewness in Figure 4.2, variance stabilization was performed using offset log transformation where a constant is added to the cases before applying log to the data with zero data values as suggested by (William & Wei, 2006). Time plot together with the ACF and PACF plots were generated from the data set that underwent transformation demonstrated in Figure 4.6.

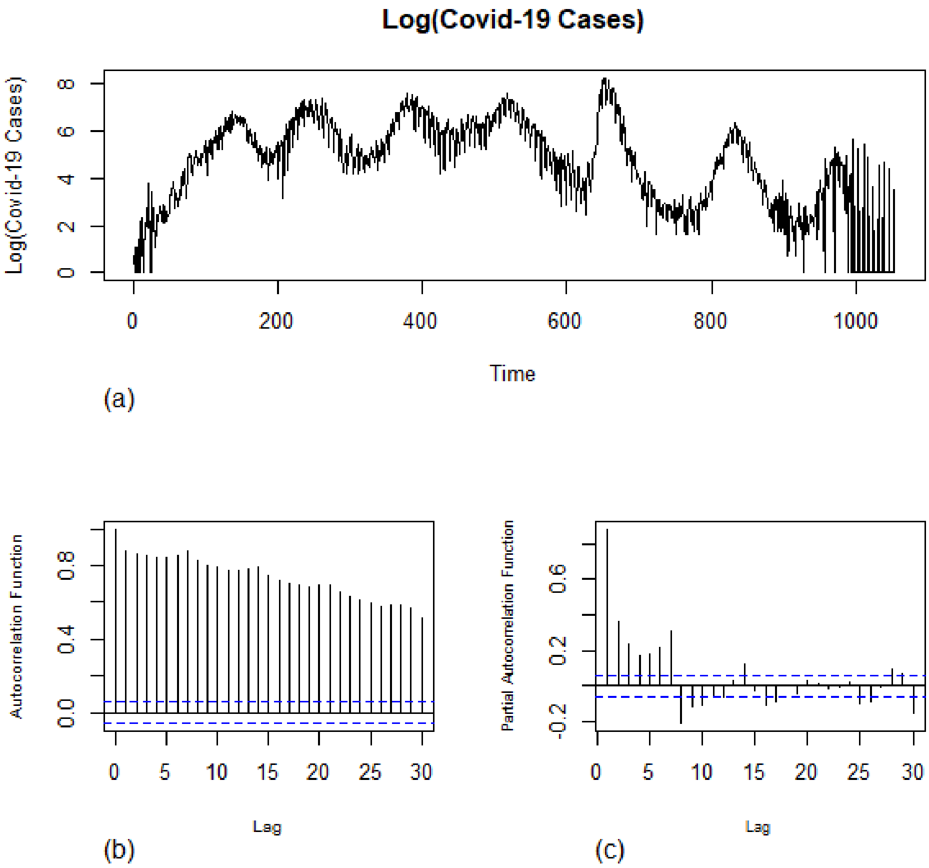

Figure 4.6: Plots of the offset log-tranformed COVID-19 cases

The slowly decaying ACF plot in Figure 4.6(a) demontrated lack of stationariy. The significant spikes at lag 7, 14, 21 and 28 indicated that weekly seasonality was present in the data. This was also observed from the weekly oscillations in the plot.

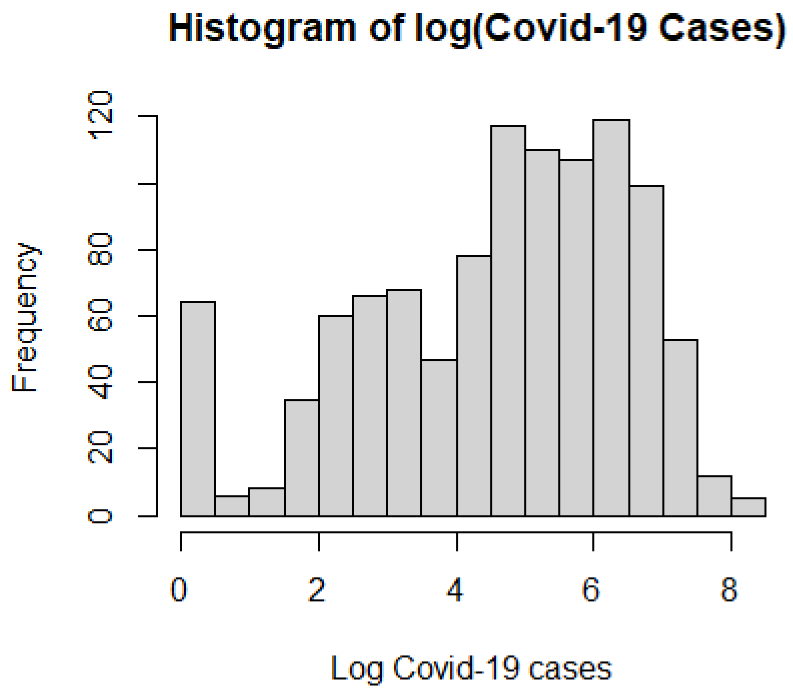

Figure 4.7: Histogram of the offset log transformed COVID-19 cases

The histogram demonstrated that the variance had been stabilized.

4.2.2 Test for stationarity

The ADF stationarity test Mushtaq (2011) was used as the means of estabilishing the stationarity of data which further determined whether differencing was needed. From the results, a $p-value$ of $.09376 > .05$ was obtained. The null hypothesis of data non-stationarity was not rejected due to lack of enough evidence at α level of significance. This led to the conclusion that data was not stationary and therefore differencing was required. Differencing was done and the mean was constant as shown in Figure 4.8(a). Figure 4.8(b) and 4.8(c) showed the ACF and PACF of the log-differenced COVID-19 cases.

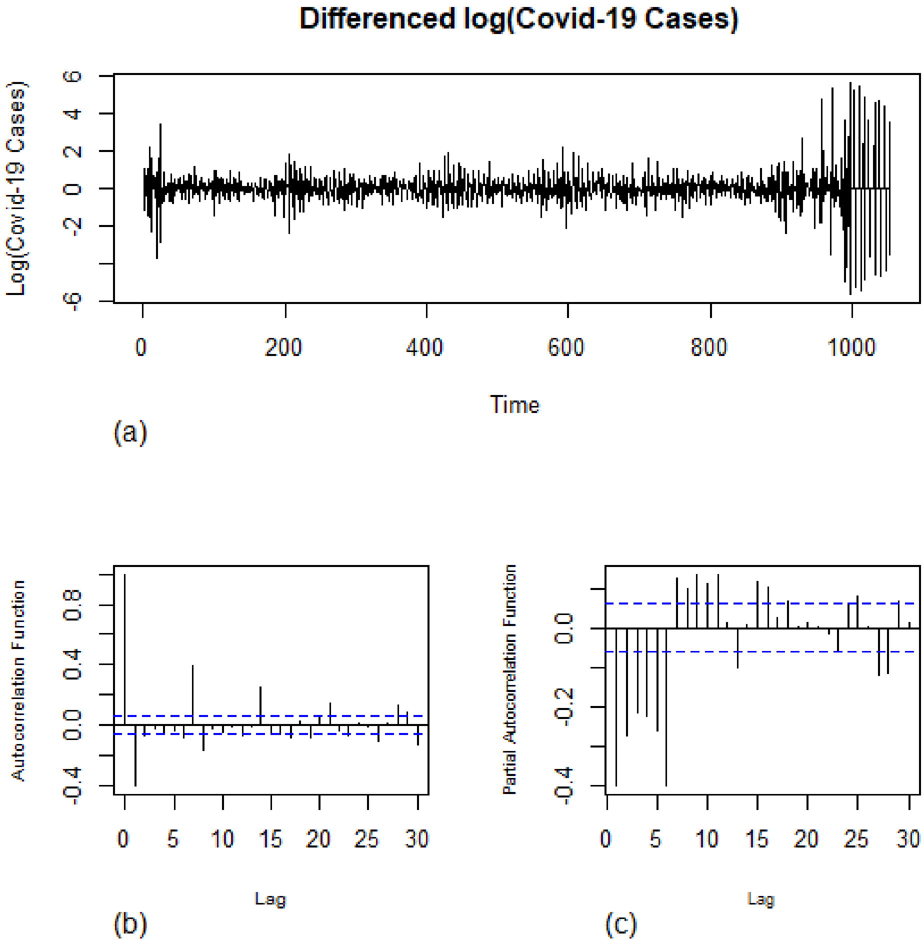

Figure 4.8: Plots of log-transformed COVID-19 cases after regular differencing

ACF plot in Figure 4.8(b) had significant spikes at lags 1, 2 but cut off at lag 3. ADF test was carried out again to evaluate the data's stationarity and the result was a $p-value$ of .01 the null hypothesis of data non-stationarity was rejected indicating that the data was stationary after regular differencing.

4.2.3 Developing the ARIMA model

According to Figure 4.8(b) and 4.8(c), the ARIMA orders were determined and the results shown in Table 4.2.

Table 4.2: Fitting the ARIMA Model

ARIMA Order	AIC	BIC	RMSE
(1,1,1)	1680.07	1694.759	0.5636
(1,1,2)	1675.17	1694.761	0.5617
(2,1,1)	1672.25	1691.837	0.5608
(2,1,2)	1605.70	1630.185	0.5416
(3,1,1)	1668.03	1692.514	0.5590
(3,1,2)	1583.18	1612.558	0.5349

During model selection, ARIMA model of order (3,1,2) was selected since it had the least AIC, BIC and least RMSE values. This results implied that the orders of the regular ARIMA model were as follows,

p - 3 (AR model order)

d - 1 (differencing order)

q - 2 (MA model order)

4.2.4 Testing the model's Adequacy

The residuals of the ARIMA (3,1,2) were tested using Box-Ljung.The resulting $p-value$ of $0.2561 > 5\%$ significance level. Therefore, the null hypothesis that the residuals had no autocorrelation was not rejected. This meant that residuals had zero autocorrelation. The ACF and PACF plots of these residuals were generated and presented in Figure 4.9 repectively.

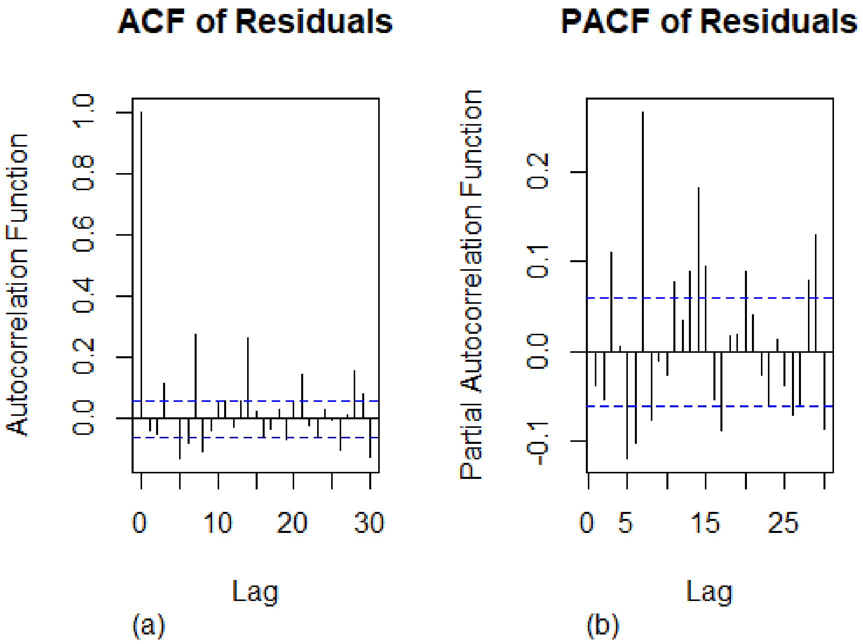

Figure 4.9: ACF and PACF of ARIMA(3,1,2) Residuals

The ACF of the residuals did not indicate perfect non Autocorrelation. This was an indication that more information could still be extracted from the data and be used in model determination (Tay, 2017). Both the ACF and PACF plots indicated that COVID-19 infections had weekly seasonality because of the significant spikes at lag 7, 14, 21 and 28. After further literature review, it was therefore suggested that data should be seasonally differenced and fit a SARIMA model. Working on the ARIMA model's forecasts is in progress.

4.3 Fitting the SARIMA model

4.3.1 Seasonal Differencing

Due to weekly seasonality from the ACF in Figure 4.9, seasonal differencing was done at lag 7. The Figure 4.10 demonstrates the time series plots generated by the seasonal differenced data.

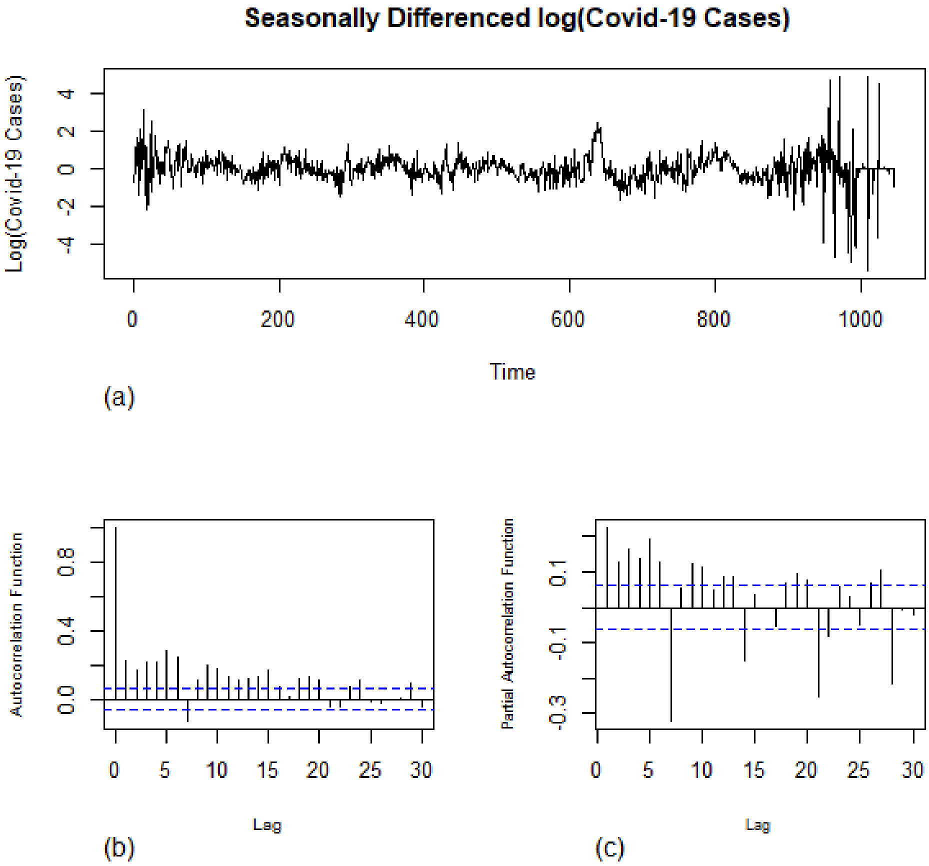

Figure 4.10: Plots of COVID-19 cases after seasonal differencing

The data had a constant mean as shown in Figure 4.10(a) hence the trend component had been removed from the COVID-19 data. The plot of the ACF also cut off which meant that the data was stationary. The ACF had significant spikes at lags 7 and 14 indicating a possible seasonal MA of order 2 or 1. The PACF also had spikes at lags 7, 14, 21 and 28 suggesting a seasonal AR of order 1,2,3 or 4. Both the ACF and PACF revealed non-seasonal MA and AR orders respectively, ranging from 0-6 each. This was because of the significant spikes in both Figure 4.10(b) and 4.10(c).

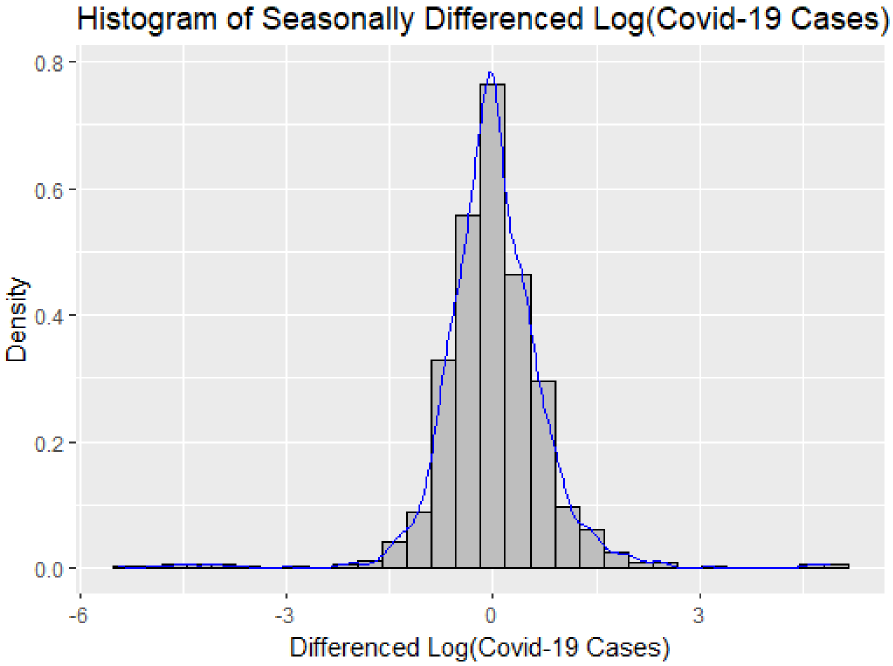

Figure 4.11: Histogram of Seasonally Differenced COVID-19 data

The histogram Figure 4.11 revealed that the offset log-transformed and differenced COVID-19 cases were approximately normally distributed. However, after seasonal differencing, data stationarity was confirmed using an ADF test. The resulting $p-value$ was 0.01, therefore the null hypothesis of data being non-stationary was rejected at 5% level of significance. These results confirmed that the data was stationary after seasonal differencing. There was therefore no need for further differencing. The data had weekly seasonality hence a seasonal ARIMA was the best model to be fitted to the data (T.-M. Choi *et al.*,2011).

4.3.2 Developing the SARIMA model

The possible model orders were first fitted to the data. The ACF and PACF plots of the offset log-transformed and differenced data generated the SARIMA model's orders. Results shown in table 4.3.

Table 4.3: Fitting the SARIMA Model

	SARIMA Order	**AIC**	**RMSE**	**BIC**
COVID-19 Cases	$(1,0,1)(1,1,1)_7$	2088.16	0.6486	2117.925
	$(1,0,1)(1,1,2)_7$	2088.33	0.6479	2118.051
	$(1,0,2)(1,1,1)_7$	2089.02	0.6482	2118.744
	$(1,0,2)(0,1,2)_7$	2094.70	0.6501	2124.423
	$(1,0,2)(2,1,1)_7$	2087.02	0.6470	2121.698
	$(1,0,1)(2,1,2)_7$	2082.50	0.6455	2117.176
	$(2,0,1)(1,1,2)_7$	2089.36	0.6477	2124.043
	$(2,0,2)(1,1,1)_7$	2092.04	0.6485	2126.716
	$(1,0,0)(1,1,1)_7$	2411.93	0.7587	2431.749
	$(2,0,2)(1,1,2)_7$	2085.52	0.6459	2131.751
	$(2,0,1)(1,1,1)_7$	2089.24	0.6483	2118.963

Information Criterions AIC and BIC guided in deciding the best fit model. SARIMA$(1,0,1)(2,1,2)_7$ was selected as the best model.

4.3.3 Testing the model's Adequacy

The residuals diagnostics of the model SARIMA$(1,0,1)(2,1,2)_7$ were tested using ACF plot, the histogram for normality and the Ljung Box test for autocorrelation (Pagan & Hall, 1983).

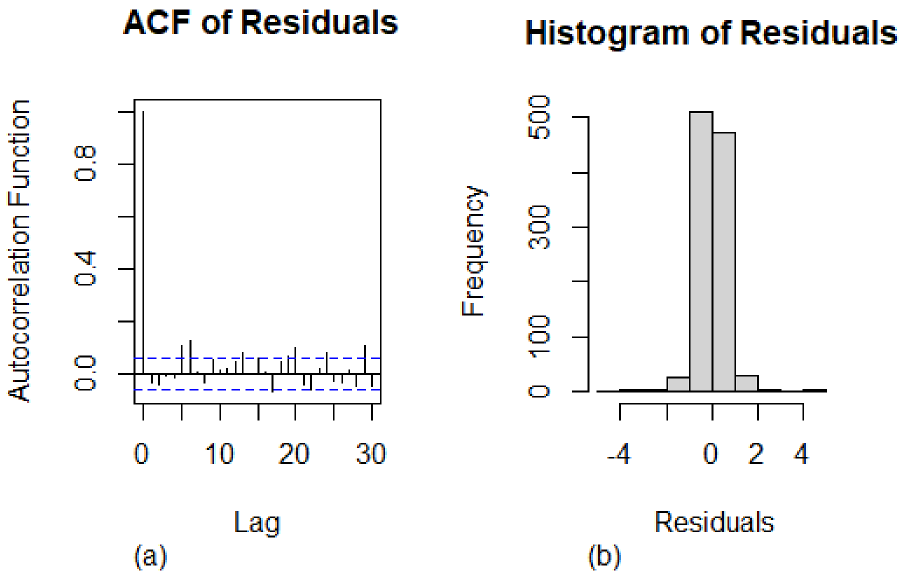

Figure 4.12: ACF and Histogram of SARIMA$(1,0,1)(2,1,2)_7$ Residuals

The histogram Figure 4.12 indicated that the residuals resembled a normal distribution. Although the ACF in Figure 4.12(a) did not indicate perfect zero autocorrelation, the autocorrelation Box-Ljung test proved that the residuals had zero autocorrelation.

The residuals were therefore independent of each other satisfying the assumption of the error terms that they are identically and independently distributed (Schielzeth *et al.*, 2020). This was deduced from the resulting $p - value$ of $.3105 > 5\%$ level of significance.

4.4 Model Validation

Tanner *et al.*, (2019) suggested the hold out technique in model validation. COVID-19 data set was split into two parts. The first part is the model fitting data set and the second part was validation data set. This method was used because of the many data points in the data and also because the data was time series data. In model fitting, 80% of the data was used. The next step involved model validation using the remaining 20% of the data. Out-of-sample prediction using test data was performed to validate the selected model according to Vrigazova (2021). The models were used in forecasting for 20% test data set. The MAE, MSE and the RMSE of the forecast errors of the different selected models were observed and compared as follows;

Table 4.4: Model Validation

	SARIMA Order	**MAE**	**MSE**	**RMSE**
COVID-19 Cases	$(1,0,1)(1,1,1)_7$	3.0860	22.7889	4.7738
	$(1,0,1)(1,1,2)_7$	3.0063	21.3473	4.6203
	$(1,0,2)(1,1,1)_7$	3.1204	23.4844	4.8461
	$(1,0,2)(0,1,2)_7$	3.2609	26.8992	5.1864
	$(1,0,2)(2,1,1)_7$	2.9882	21.0197	4.5847
	$(1,0,1)(2,1,2)_7$	2.9867	20.9905	4.5815
	$(2,0,1)(1,1,2)_7$	3.0323	21.7943	4.6684
	$(2,0,2)(1,1,1)_7$	3.0891	22.8524	4.7804
	$(1,0,0)(1,1,1)_7$	5.4582	134.4129	11.5937
	$(2,0,2)(1,1,2)_7$	3.0025	21.2827	4.6133
	$(2,0,1)(1,1,1)_7$	3.1140	23.3495	4.8321

From the Table 4.4, the model SARIMA(1,0,1)(2,1,2)$_7$ had the least validation statistics. This lead to the conclusion that it was the model of best fit. The RMSE of the selected model was then normalized for confirmation of validation results due to the seasonal component (Shcherbakov *et al.*, 2013).

$$\begin{aligned}\text{Normalized RMSE} &= \frac{RMSE}{\text{Max actual value-Min actual value}}\\ \text{Normalized RMSE} &= \frac{4.5815}{(3749-0)}\\ &= 0.0012 \end{aligned} \tag{4.4.1}$$

SARIMA(1,0,1)(2,1,2)$_7$ model could therefore be used for prediction because its normalized RMSE was closer to 0 than 1. The NRMSE value ranges between 1 and 0. The model SARIMA(1,0,1)(2,1,2)$_7$ forecasts for the test set are shown in the plot 4.13.

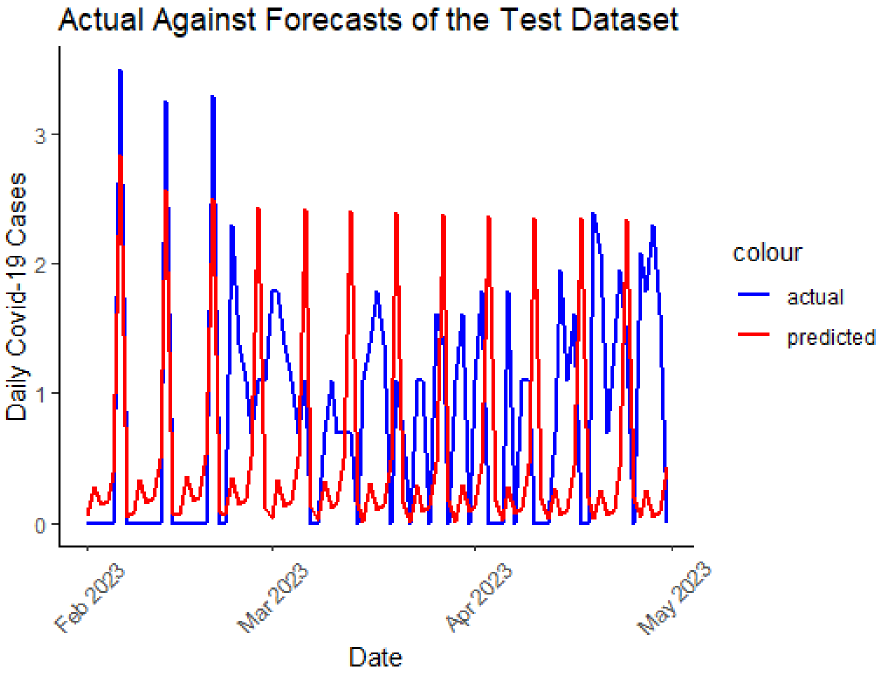

Figure 4.13: Plot of 20%() test data-set forecasts of daily COVID-19 cases using SARIMA model against actual COVID-19 data

SARIMA Model Parameters

Table 4.5: SARIMA Model coefficients, AIC and BIC values

	AR	SAR	MA	SMA	AIC	BIC
COVID-19 Cases	0.9868	-0.3833	-0.7527	-0.2711		
		0.2350		-0.5538	**2082.5**	**2117.18**

The model SARIMA(1,0,1)(2,1,2)$_7$ is written as

$$\begin{aligned}(1-0.9868B)(1+0.3833B^7-0.2350B^7)(1-B^7)X_t &= \\ (1-0.7527B)(1-0.2711B^7-0.5538B^7)e_t\end{aligned} \tag{4.4.2}$$

Therefore,

$$X_t = \frac{(1-0.7527B)(1-0.2711B^7-0.5538B^7)e_t}{(1-0.9868B)(1+0.3833B^7-0.2350B^7)(1-B^7)} \tag{4.4.3}$$

4.5 Forecasting

The model SARIMA(1,0,1)(2,1,2)$_7$ was then used to forecast into the future for 90 days. Table 4.6 are the forecasts of COVID-19 cases.

Table 4.6: Forecasts

Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
1.5178199	0.67220395	2.363436	0.22456190	2.811078
0.3133367	-0.55133163	1.178005	-1.00905940	1.635733
1.6022156	0.71929642	2.485135	0.25190722	2.952524
1.6915947	0.79116922	2.592020	0.31451278	3.068677
1.4407694	0.52353191	2.358007	0.03797573	2.843563
1.3534834	0.42008315	2.286884	-0.07402911	2.780996

Continued on next page

Table 4.6 – *Continued from previous page*

Point Forecast	**Lo 80**	**Hi 80**	**Lo 95**	**Hi 95**
0.9084360	-0.04051827	1.857390	-0.54286434	2.359736
1.7681195	0.70585337	2.830386	0.14352363	3.392715
0.4141473	-0.67342668	1.501721	-1.24915356	2.077448
1.3052686	0.19348051	2.417057	-0.39506457	3.005602
1.5374089	0.40241959	2.672398	-0.19840748	3.273225
1.4986390	0.34139003	2.655888	-0.27122055	3.268498
1.2994739	0.12084378	2.478104	-0.50308531	3.102033
0.9171174	-0.28207205	2.116307	-0.91688457	2.751119
1.7976871	0.50647359	3.088901	-0.17705358	3.772428
0.4951323	-0.82440235	1.814667	-1.52292179	2.513186
1.2039773	-0.14270197	2.550657	-0.85559094	3.263546
1.5248207	0.15209117	2.897550	-0.57458793	3.624229
1.2867636	-0.11099439	2.684522	-0.85092276	3.424450
1.2437428	-0.17808706	2.665573	-0.93075833	3.418244
1.1504934	-0.29451031	2.595497	-1.05944909	3.360436
1.8653110	0.37060222	3.360020	-0.42064879	4.151271
0.5232007	-0.99740778	2.043809	-1.80236931	2.848771
1.1395974	-0.40594218	2.685137	-1.22410143	3.503296
1.4926093	-0.07695006	3.062169	-0.90782457	3.893043
1.3155331	-0.27718653	2.908253	-1.12032139	3.751388
1.2382152	-0.37685275	2.853283	-1.23181809	3.708249
1.1491089	-0.48753869	2.785756	-1.35392758	3.652145
1.8756898	0.19492814	3.556451	-0.69481333	4.446193
0.5479720	-1.15678246	2.252727	-2.05922497	3.155169
1.1223279	-0.60558879	2.850245	-1.52029263	3.764948

Continued on next page

Table 4.6 – *Continued from previous page*

Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
1.4962204	-0.25407119	3.246512	-1.18061961	4.173060
1.2667413	-0.50517789	3.038660	-1.44317524	3.976658
1.2298422	-0.56299393	3.022678	-1.51206405	3.971748
1.2129321	-0.60014441	3.026009	-1.55992912	3.985793
1.8974695	0.05000842	3.744931	-0.92797839	4.722917
0.5592152	-1.30981290	2.428243	-2.29921661	3.417647
1.1120306	-0.77786957	3.001931	-1.77832227	4.002383
1.4930957	-0.41701364	3.403205	-1.42816444	4.414356
1.2809217	-0.64876396	3.210607	-1.67027785	4.232121
1.2337704	-0.71488665	3.182427	-1.74644337	4.213984
1.2149736	-0.75207595	3.182023	-1.79336906	4.223316
1.9040774	-0.09432764	3.902482	-1.15221935	4.960374
0.5697954	-1.44830749	2.587898	-2.51662665	3.656217
1.1128854	-0.92431387	3.150085	-2.00274203	4.228513
1.4988449	-0.55687463	3.554564	-1.64510681	4.642797
1.2728060	-0.80088142	3.346493	-1.89862523	4.444237
1.2361372	-0.85498814	3.327263	-1.96196303	4.434237
1.2354506	-0.87260375	3.343505	-1.98854031	4.459441
1.9137927	-0.22192767	4.049513	-1.35250973	5.180095
0.5766653	-1.57704803	2.730379	-2.71715501	3.870486
1.1146989	-1.05648211	3.285880	-2.20583589	4.435234
1.5075829	-0.80270558	3.817871	-2.02569844	5.040864
1.2504921	-1.20100078	3.701985	-2.49874273	4.999727
1.2525801	-1.21236129	3.717521	-2.51722244	5.022383
1.9301055	-0.55703500	4.417246	-1.87364763	5.733859

Continued on next page

Table 4.6 – *Continued from previous page*

Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
0.5935048	-1.90805044	3.095060	-3.23229379	4.419303
1.1274817	-1.38810312	3.643066	-2.71977326	4.974737
1.5166517	-1.01259021	4.045894	-2.35149000	5.384793
1.2920446	-1.25049424	3.834583	-2.59643297	5.180522
1.2548181	-1.30066904	3.810305	-2.65346222	5.163098
1.2580287	-1.31006933	3.826127	-2.66953833	5.185596
1.9347990	-0.65417598	4.523774	-2.02469655	5.894295
0.5979528	-2.00456629	3.200472	-3.38225670	4.578162
1.1316119	-1.48409666	3.747320	-2.86876913	5.131993
1.5208141	-1.10774055	4.149369	-2.49921331	5.540841
1.2965435	-1.34452469	3.937612	-2.74262171	5.335709
1.2590288	-1.39423072	3.912288	-2.79878149	5.316839
1.2620282	-1.40311062	3.927167	-2.81394988	5.338006
1.9389090	-0.74601758	4.623836	-2.16733185	6.045150
0.6021045	-2.09561413	3.299823	-3.52370012	4.727909
1.1356031	-1.57457833	3.845784	-3.00926170	5.280468
1.5248294	-1.19749560	4.247154	-2.63860739	5.688266
1.3002582	-1.43390104	4.034417	-2.88127750	5.481794
1.2628927	-1.48280073	4.008586	-2.93628304	5.462068

The plot of COVID-19 cases and their forecasts given by Table 4.6 from 1^{st} May, 2023 to 29^{th} July, 2023.

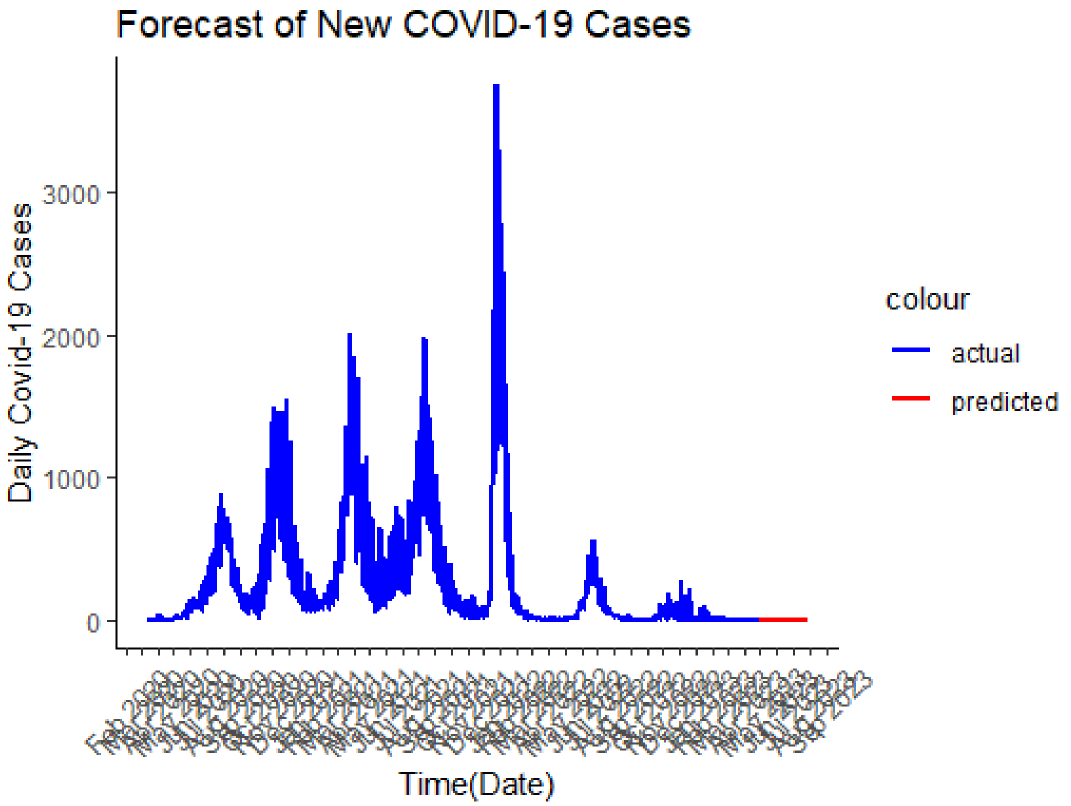

Figure 4.14: Plot of 90 days forecasts of daily COVID-19 cases using SARIMA model

The COVID-19 forecasts in Table 4.6 were plotted with the historical dataset as shown in Figure 4.14. The forecasts cases will continue decreasing but not to zero. The forecasts revealed a donward trend in the number of COVID-19 cases.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

The summary of this research and conclusions drawn based on the findings are mentioned in this chapter. The challenges met during the research process have also been mentioned here and the recommendations have been made.

5.2 Summary

After the emergence of the COVID-19 pandemic on March 2020 , alot of research has been done about its trend using time series analysis approach. The ARIMA and SARIMA models however, had not been fitted to the long COVID-19 data set in Kenya. This research therefore fitted ARIMA and SARIMA models to the long Kenyan daily COVID-19 dataset.

Given that the data was not stable, an offset log transformation was used to stabilize the variance. This is when a constant(offset) is first added to the data values before finding their logarithm. The data also had a trend component which was removed by first differencing. The possible ARIMA model orders were then determined from ACF and PACF plots of the log-differenced data. The models were fitted to the data and ARIMA(3,1,2) was selected as the best with the least AIC and BIC values. The residuals of the selected ARIMA(3,1,2) revealed that the data contained a seasonal component.

The seasonal component was removed by seasonal differencing. The possible SARIMA model orders were determined from ACF and PACF plots of the log-differenced data. After fitting all the possible SARIMA model orders the model with the least AIC, BIC and RMSE was picked to proceed to the next stage. The generation of parameter estimates was done using MLE approach and the models were then used to forecast for the test data set.

The selected model SARIMA(1,0,1)(2,1,2)$_7$ was taken through residual diagnostic checking. The model's accuracy was tested by using forecasts in the test data set to obtain forecast errors, and the model with the lowest validation statistics(errors) was selected. Further, the normalized RMSE was also used to validate the selected model bacause there was weekly seasonality in the data.

The selected model was then used to forecast for 90 days into the future.

5.3 Conclusion

COVID-19 has been a pandemic for almost three years now therefore there was a great desire to know if the pandemic would eventually end or not. ARIMA and SARIMA models were fitted onto the data.

Due to variance instability, offset log-transformation was performed on COVID-19 training dataset for stability. The generated ACF and PACF plots then suggested that the data was not stationary and it also had a weekly seasonality which is categorized as weak seasonality. The data was therefore first differenced to achieve stationarity. ADF test was done and it confirmed that the data was stationary after first differencing. The ACF and PACF plots of the log-differenced COVID-19 data were used to determine the possible ARIMA model orders. ARIMA(3,1,2) emerged as the model with the least AIC, BIC and RMSE values amongst all possible models that were fitted to the data.

ACF and PACF plots of the model's residuals were then plotted. Both plots revealed that the data indeed had a seasonal component. This suggested seasonal differencing and a possibility of a SARIMA model.

Seasonal differecing was perfomed and the results demonstrated by Figure 4.10. The possible SARIMA model orders were then determined from ACF and PACF plots of the log-differenced data. From the set of possible models, SARIMA (1,0,1)(2,1,2)$_7$ was the best model because it had the least AIC, BIC and RMSE values. The fitted models parameters were estimated using the Maximum Likelihood Estimation method. The model was then taken through the validation step. As the ACF of the residuals revealed that they had zero autocorrelation, the Box-Ljung test confirmed that the residuals were white noise with a $p - value$ of .01. According to the histogram of the residuals, the residuals followed a normal distribution. The forecast errors of the all the possible models were compared using MAE, MSE and RMSE. SARIMA(1,0,1)(2,1,2)$_7$ model had the least validation statistics as, MAE = 2.9867, MSE = 20.9905 and RMSE = 4.5815. The RMSE value was then normalized as part of model validation. The resulting NRMSE= 0.0012 revealed that the model was a good fit to the data and therefore could be used for future prediction. According to the forecast errors obtained from comparing the actual COVID-19 cases the models forecasts can be used and trusted. The model was found to be accurate with only 4.58$\%$ error.

The model was then fitted to total COVID-19 cases data for forecasting. The results indicated that the reported COVID-19 recorded cases would decrease as shown in Figure 4.14. Although COVID-19 cases reduced, the cases did not drop to zero hence the virus can no longer be termed as a pandemic but also cannot be ignored.

The findings of this study may be helpful to researchers, Kenyan government and the stakeholders to examine the possible disease burden during the pandemic. The pandemic affected the country's achievement of SDG's and Big 4 agenda on the set time frame. This model has proven to be adequate and sufficient. It can therefore be used for forecasting as more data becomes available or in case of any other future pandemic diseases. Its applicability can be used to make predictions of cases. People are therefore advised to keep watch for such contagious diseases

5.4 Recommendations

This research recommends for more exploration in time series modelling of daily COVID-19 cases as the data points increases. A comparative study on Bayesian SARIMA and SARIMA model can be done in modelling daily COVID-19 in Kenya. Due to the lifting and relaxation of restrictions by the Kenyan government in March 2022, one can try and fit different models before and after the restrictions due to possible change in probabilistic structures of the data. The data having been collected on a daily basis was found to have weekly seasonality. This was found in the ACF of ordinary differenced data. Having weekly seasonality would also imply presence of monthly seasonality. With the presence of two or more types of seasonality, one can fit the BATS and TBATS models to the data.

References

Al Khames Aga, Q. A., Alkhaffaf, W. H., Hatem, T. H., Nassir, K. F., Batineh, Y., Dahham, A. T., … Traqchi, M. (2021). Safety of covid-19 vaccines. *Journal of medical virology*, *93*(12), 6588–6594.

Anderson, T. W. (2011). *The statistical analysis of time series*. John Wiley & Sons.

Baglaeva, E., Sergeev, A., Shichkin, A., & Buevich, A. (2020). The effect of splitting of raw data into training and test subsets on the accuracy of predicting spatial distribution by a multilayer perceptron. *Mathematical Geosciences*, *52*, 111–121.

Bee Dagum, E., & Bianconcini, S. (2016). Time series components. In *Seasonal adjustment methods and real time trend-cycle estimation* (pp. 29–57). Springer.

Bhangu, K. S., Sandhu, J. K., & Sapra, L. (2021). Time series analysis of covid-19 cases. *World Journal of Engineering*, *19*(1), 40–48.

Brockwell, P. J., & Davis, R. A. (2009). *Time series: theory and methods*. Springer science & business media.

Cavanaugh, J. E., & Neath, A. A. (2019). The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, *11*(3), e1460.

Chai, T., & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae). *Geoscientific model development discussions*, *7*(1), 1525–1534.

Chakraborty, T., Ghosh, I., Mahajan, T., & Arora, T. (2022). Nowcasting of covid-19 confirmed cases: Foundations, trends, and challenges. *Modeling, Control and Drug Development for COVID-19 Outbreak Prevention*, 1023–1064.

Choi, B. (2012). *Arma model identification*. Springer Science & Business Media.

Choi, T.-M., Yu, Y., & Au, K.-F. (2011). A hybrid sarima wavelet transform method for sales forecasting. *Decision Support Systems*, *51*(1), 130–140.

Coghlan, A. (2018). A little book of r for time series. *Release 0.2*, *75*.

Coşkun, H., Yıldırım, N., & Gündüz, S. (2021). The spread of covid-19 virus through population density and wind in turkey cities. *Science of the Total Environment*, *751*, 141663.

Dehesh, T., Mardani-Fard, H. A., & Dehesh, P. (2020). Forecasting of covid-19 confirmed cases in different countries with arima models. *MedRxiv*, 2020–03.

Devi, B. U., Sundar, D., & Alli, P. (2013). An effective time series analysis for stock trend prediction using arima model for nifty midcap-50. *International Journal of Data Mining & Knowledge Management Process*, *3*(1), 65.

Dickey, D. A. (2015). Stationarity issues in time series models. *SAS Users Group International*, *30*.

Dritsakis, N., & Klazoglou, P. (2018). Forecasting unemployment rates in usa using box-jenkins methodology. *International Journal of Economics and Financial Issues*, *8*(1), 9.

Duong, D. (2021). *Alpha, beta, delta, gamma: What's important to know about sars-cov-2 variants of concern?* Can Med Assoc.

Ebhuoma, O., Gebreslasie, M., & Magubane, L. (2018). A seasonal autoregressive integrated moving average (sarima) forecasting model to predict monthly malaria cases in kwazulu-natal, south africa. *South African medical journal*, *108*(7).

Feroze, N. (2020). Forecasting the patterns of covid-19 and causal impacts of lockdown in top five affected countries using bayesian structural time series models. *Chaos, Solitons and Fractals*, *140*, 110196.

Gebretensae, Y. A., & Asmelash, D. (2021). Trend analysis and forecasting the spread of covid-19 pandemic in ethiopia using box–jenkins modeling procedure. *International journal of general medicine*, 1485–1498.

Ho, S.-L., Xie, M., & Goh, T. N. (2002). A comparative study of neural network and box-jenkins arima modeling in time series prediction. *Computers & Industrial Engineering*, *42*(2-4), 371–375.

Hu, W., Tong, S., Mengersen, K., & Connell, D. (2007). Weather variability and the incidence of cryptosporidiosis: comparison of time series poisson regression and sarima models. *Annals of epidemiology*, *17*(9), 679–688.

Jayaweera, M., Perera, H., Gunawardana, B., & Manatunge, J. (2020). Transmission of covid-19 virus by droplets and aerosols: A critical review on the unresolved dichotomy. *Environmental research*, *188*, 109819.

Khan, F. M., & Gupta, R. (2020). Arima and nar based prediction model for time series analysis of covid-19 cases in india. *Journal of Safety Science and Resilience*, *1*(1), 12–18.

Kiarie, J., Mwalili, S., & Mbogo, R. (2022). Forecasting the spread of the covid-19 pandemic in kenya using seir and arima models. *Infectious Disease Modelling*, *7*(2), 179–188.

Koyuncu, K., Tavacioğlu, L., Gökmen, N., & Arican, U. Ç. (2021). Forecasting covid-19 impact on rwi/isl container throughput index by using sarima models. *Maritime Policy & Management*, *48*(8), 1096–1108.

LeBaron, B., & Weigend, A. S. (1998). A bootstrap evaluation of the effect of data splitting on financial time series. *IEEE Transactions on Neural Networks*, *9*(1), 213–220.

Lima, C. M. A. d. O. (2020). *Information about the new coronavirus disease (covid-19)* (Vol. 53). SciELO Brasil.

Márquez, F. P. G., Pedregal, D. J., & Roberts, C. (2015). New methods for the condition monitoring of level crossings. *International Journal of Systems Science*, *46*(5), 878–884.

Maurice, W., Kitavi, D. M., Mugo, D. M., & Atitwa, E. B. (2021). Covid-19 prediction in kenya using the arima model. *International Journal of Electrical Engineering and Technology (DJEET)*, *12*(8).

Moh'dMussa, A., & Saxena, K. (2018). Trend analysis and forecasting of performance of students in mathematics in certificate secondary education examination in zanzibar: Arima modelling approach.

Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather*, *16*(1), 69–88.

Moskovitch, R., & Shahar, Y. (2015). Classification-driven temporal discretization of multivariate time series. *Data Mining and Knowledge Discovery*, *29*, 871–913.

Mushtaq, R. (2011). Augmented dickey fuller test.

Naqvi, S. N. Z., Yfantidou, S., & Zimányi, E. (2017). Time series databases and influxdb. *Studienarbeit, Université Libre de Bruxelles*, *12*.

Neath, A. A., & Cavanaugh, J. E. (2012). The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 199–203.

Odhiambo, J. O., Okungu, J. O., & Mutuura, C. G. (2020). Stochastic modeling and prediction of the covid-19 spread in kenya. *Engineering Mathematics*, *4*(2), 31.

Pagan, A. R., & Hall, A. D. (1983). Diagnostic tests as residual analysis. *Econometric Reviews*, *2*(2), 159–218.

Perone, G. (2021). Comparison of arima, ets, nnar, tbats and hybrid models to forecast the second wave of covid-19 hospitalizations in italy. *The European Journal of Health Economics*, 1–24.

Perone, G. (2022). Using the sarima model to forecast the fourth global wave of cumulative deaths from covid-19: Evidence from 12 hard-hit big countries. *Econometrics*, *10*(2), 18.

Pesaran, M. H. (2007). A simple panel unit root test in the presence of cross-section dependence. *Journal of applied econometrics*, *22*(2), 265–312.

Prabhakaran, S. (2019). Arima model–complete guide to time series forecasting in python. *Machine Learning Plus*, *18*.

Razali, N. M., Wah, Y. B., et al. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, *2*(1), 21–33.

Samson, T. K., Ogunlaran, O. M., & Raimi, M. O. (2020). A predictive model for confirmed cases of covid-19 in nigeria. *TK Samson, OM Ogunlaran, OM Raimi (2020)*, 1–10.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., … Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in ecology and evolution*, *11*(9), 1141–1152.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Serra, R., & Rodríguez, A. C. (2012). The ljung-box test as a performance indicator for vircs. In *International symposium on electromagnetic compatibility-emc europe* (pp. 1–6).

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., Kamaev, V. A., et al. (2013). A survey of forecast error measures. *World applied sciences journal*, *24*(24), 171–176.

So, E. C. (2013). A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? *Journal of Financial Economics*, *108*(3), 615–640.

So, I., & AM, J. (2009). Comparison of criteria for estimating the order of autoregressive process: A monte carlo approach. *European Journal of Scientific Research*, *30*(3), 409–416.

Struyf, T., Deeks, J. J., Dinnes, J., Takwoingi, Y., Davenport, C., Leeflang, M. M., . . . others (2022). Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has covid-19. *Cochrane Database of Systematic Reviews*(5).

Swain, P. K., Tripathy, M. R., Jena, D., Fenta, H. M., & Zike, D. T. (2020). Modelling and forecasting of covid-19 cases in odisha and india. *Demography India*, *9*(Special Issue), 66–75.

Tan, C. V., Singh, S., Lai, C. H., Zamri, A. S. S. M., Dass, S. C., Aris, T. B., . . . Gill, B. S. (2022). Forecasting covid-19 case trends using sarima models during the third wave of covid-19 in malaysia. *International journal of environmental research and public health*, *19*(3), 1504.

Tanner, E. M., Bornehag, C.-G., & Gennings, C. (2019). Repeated holdout validation for weighted quantile sum regression. *MethodsX*, *6*, 2855–2860.

Tay, D. (2017). Time series analysis of discourse: A case study of metaphor in psychotherapy sessions. *Discourse studies*, *19*(6), 694–710.

Valipour, M. (2015). Long-term runoff study using sarima and arima models in the united states. *Meteorological Applications*, *22*(3), 592–598.

Vartanian, T. P. (2010). *Secondary data analysis*. Oxford University Press.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological methods*, *17*(2), 228.

Vrigazova, B. (2021). The proportion for splitting data into training and test set for the bootstrap in classification problems. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, *12*(1), 228–242.

Wambua, S., Malla, L., Mbevi, G., Kandiah, J., Nwosu, A.-P., Tuti, T., … Okiro, E. A. (2022). Quantifying the indirect impact of covid-19 pandemic on utilisation of outpatient and immunisation services in kenya: a longitudinal study using interrupted time series analysis. *BMJ open*, *12*(3), e055815.

William, W., & Wei, S. (2006). Time series analysis: univariate and multivariate methods. *USA, Pearson Addison Wesley, Segunda edicion. Cap*, *10*, 212–235.

Zhang, M. (2018). Time series: Autoregressive models ar, ma, arma, arima. *University of Pittsburgh*.

Appendix

R Code

```
library(lubridate)
library(timeSeries)
library(ggplot2)
library(fBasics)
library(tseries)
library(forecast)
library(dplyr)
library(readxl)
library(stats)
library(R2jags)
library(rjags)
library(runjags)
library(coda)
library(mosaic)
library(fpp2)
library(urca)
library(fpp3)
library(ggfortify)
library(TSstudio)
library(MASS)
#library(tidyverse)
Covid_data <-
 ↪  read_excel("D:/Project/data/CovidData1.xlsx")
# Convert date variable to Date class
```

```
Covid_data$Date <- as.Date(Covid_data$Date, format =
 ↪  "%Y-%m-%d")
#Training set
mydata <- subset(Covid_data, Date<= "2023-01-31")
#test set
testdata <- subset(Covid_data, Date>="2023-02-01")
convert <- ts(testdata$New_cases)

# Converting the data to time series
Df <- ts(mydata$New_cases,start=c(2020,3),frequency=365)
data_subset <- window(mydata$New_cases, start = "", end =
 ↪  c(2023,1))

ggplot(Covid_data,aes(x=Date))+
geom_line(aes(y=New_cases),color="blue")+
labs(y="New_cases",x="Date")+ggtitle("Covid-19
 ↪  Cases")+theme_classic()

%#Plot the histogram
Histogram <- hist(data_subset)

%#ACF and PACF plots
par(mfrow = c(1,2))
ACF <- acf(data_subset,xlab = "Lag", ylab = "ACF",
main = "COVID-19 cases")

PACF <- pacf(data_subset, xlab = "Lag", ylab = "Partial
 ↪  ACF",
                              main = "COVID-19 cases")

%#Descriptives for new_cases
basicStats(data_subset)
shapiro.test(data_subset)

%# Conduct the augmented Dickey-Fuller test for
 ↪  stationarity
adf.test(data_subset, alternative=c("explosive"),k=1)

constant <-1
testdataset <- log(convert+ constant)
testdataset
Logdata <- log(data_subset+constant)
%#test for normality
shapiro.test(Logdata)
acf(Logdata ,xlab = "Lag", ylab = "ACF",
main = "Log COVID-19 cases")
pacf(Logdata, xlab = "Lag", ylab = "Partial ACF",
main = "Log COVID-19 cases")
%# Plot a time series plot of the logtransformed data
ggtsdisplay(Logdata)

%#Plotting a histogram
hist(Logdata,xlab="COVID-19 cases",main="Histogram of
 ↪  COVID-19 cases")

%#Testing for stationarity: Found stationary
```

```
adf.test(Logdata)
%#First Differencing
DifferencedData1 <- diff(Logdata)

%# Plot a time series plot of the logdifferenced data
gtsdisplay(DifferencedData1)

%# acf and pacf of ts data
acf(DifferencedData1 , xlab = "Lag", ylab = "ACF",
main = "Differenced COVID-19 cases")
pacf(DifferencedData1, xlab = "Lag", ylab = "Partial
 ↪  ACF",
                                main = "Differenced
                                 ↪  COVID-19 cases")
hist(DifferencedData1,xlab = "Differenced COVID-19
 ↪  cases",
                                main = "Histogram of
                                 ↪  Differenced COVID-19
                                 ↪  cases")

Model1 <- arima(Logdata, order=c(1,1,1))
summary(Model1)
BIC(Model1)
acf(resid(Model1), main=" Residuals")

Model12 <- arima(Logdata, order=c(1,1,2))
summary(Model2)
BIC(Model2)
```

```
acf(resid(Model2), main=" Residuals")

Model3 <- arima(Logdata, order=c(2,1,1))
summary(Model3)
BIC(Model3)
acf(resid(Model3), main=" Residuals")
Model4 <- arima(Logdata, order=c(2,1,2))
summary(Model4)
BIC(Model4)
acf(resid(Model4), main=" Residuals")

Model5 <- arima(Logdata, order=c(3,1,1))
summary(Model5)
BIC(Model5)
acf(resid(Model5), main=" Residuals")

Model6 <- arima(Logdata, order=c(3,1,2))
summary(Model6)
BIC(Model6)
acf(resid(Model6), main=" Residuals")

%#Seasonal Differencing
DifferencedData <- diff(Logdata,lag=7)

%# Plot a time series plot of the logdifferenced data
ggtsdisplay(DifferencedData)

%# acf and pacf of ts data
```

```
acf(DifferencedData , xlab = "Lag", ylab = "ACF",
main = "Differenced COVID-19 cases")
pacf(DifferencedData, xlab = "Lag", ylab = "Partial ACF",
main = "Differenced COVID-19 cases")

hist(DifferencedData,xlab = "Differenced COVID-19 cases",
main = "Histogram of Differenced COVID-19 cases")
%%####--------------------------------------------------------------
%%#Fitting the SARIMA models
ModelA <- arima(Logdata, order=c(1,0,1), seasonal =
 ↪  list(order=c(1,1,1), period=7),include.mean = FALSE
                        , method="ML")
summary(ModelA)
ModelB <- arima(Logdata, order=c(1,0,1), seasonal =
 ↪  list(order=c(1,1,2), period=7),include.mean = FALSE,
 ↪  method="ML")
ModelD <- arima(Logdata, order=c(1,0,2), seasonal =
 ↪  list(order=c(0,1,2), period=7),include.mean = FALSE,
 ↪  method="ML")
ModelE <- arima(Logdata, order=c(1,0,2), seasonal =
 ↪  list(order=c(2,1,1), period=7),include.mean = FALSE,
 ↪  method="ML")
ModelF <- arima(Logdata, order=c(1,0,1), seasonal =
 ↪  list(order=c(2,1,2), period=7),include.mean = FALSE,
 ↪  method="ML")
ModelG <- arima(Logdata, order=c(2,0,1), seasonal =
 ↪  list(order=c(1,1,2), period=7),include.mean = FALSE,
 ↪  method="ML")
```

```
ModelH <- arima(Logdata, order=c(2,0,2), seasonal =
 ↪  list(order=c(1,1,1), period=7),include.mean = FALSE,
 ↪  method="ML")
ModelI <- arima(Logdata, order=c(1,0,0), seasonal =
 ↪  list(order=c(1,1,1), period=7),include.mean = FALSE,
 ↪  method="ML")
ModelJ <- arima(Logdata, order=c(2,0,2), seasonal =
 ↪  list(order=c(1,1,2), period=7),include.mean = FALSE,
 ↪  method="ML")

#Fitting the selected model S-ARIMA(1,0,1)(2,1,2)7
Transformedcases <- log(Covid_data$New_cases +
 ↪  constant)
ModelFit <- arima(Transformedcases, order=c(1,0,1),
 ↪  seasonal = list(order=c(2,1,2),
 ↪  period=7),include.mean = FALSE, method="ML")
summary(ModelFit)
resid(ModelFit)
BIC(ModelFit)
acf(resid(ModelFit), main=" Residuals")
Box.test(resid(ModelFit),type="Ljung")
forecastsModelFit<-forecast(ModelFit,90)
```