MODELLING COVID-19 PANDEMIC IN KENYA

MOURICE WANYONYI
B531/1313/2019

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN STATISTICS OF THE UNIVERSITY OF EMBU

SEPTEMBER, 2021

DECLARATION

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

Signature …............................... Date.....................................

Mourice Wanyonyi
Department of Mathematics and Statistics
B531/1313/2019

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

Signature ………………………. Date …………………………

Dr. Dominic Makaa Kitavi
Department of Mathematics and Statistics
University of Embu.

Signature ………………………... Date……………………………

Dr. David Muchangi Mugo
Department of Computing and Information Technology
University of Embu.

Signature…………………………. Date……………………………

Dr. Edwin Benson Atitwa
Department of Mathematics and Statistics
University of Embu.

DEDICATION

I dedicate this research project to my lovely wife Metrine and son Damian. A special dedication also goes to my lovely parents and guardians for their overwhelming financial support and encouragement.

ACKNOWLEDGEMENT

I would like to thank the Kenya Ministry of Health for publishing the COVID-19 data on their official website. My sincere gratitude goes to my supervisors, Dr. Dominic M. Kitavi, Dr. David M. Mugo and Dr. Edwin B. Atitwa and whose advice, encouragement and time enabled me to accomplish this project work. Finally, my gratitude goes to the University of Embu Administration for awarding me the scholarship that enabled me pursue my Master's Degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION AND ACRONYMS

ACF	Autocorrelation Function
AIC	Akaike Information Criteria
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
BIC	Bayesian Information Criteria
CDC	Centre for Disease Control and Prevention
COVID-19	Coronavirus Disease 2019
LSE	Least Square Estimator
MA	Moving Average
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLE	Maximum Likelihood Estimator
MOH	Ministry of Health
MSE	Mean Square Error
PACF	Partial Autocorrelation Function
PPE	Personal Protective Equipment
RMSE	Root Mean Square Error
SARS-COV-2	Severe Acute Respiratory Syndrome Coronavirus 2
WHO	World Health Organization
WN	White Noise

LIST OF SYMBOLS

φ	varphi
ϕ	Phi variant
σ	sigma
β	beta
ξ	xi
τ	tau
κ	kappa
γ	gamma
ε	epsilon
ρ	rho
μ	mu
∇	nabla
θ	theta
ψ	psi
ω	omega
α	alpha
η	eta
ζ	zeta
π	pi
Λ	lambda
Φ	phi

ABSTRACT

Due to increased COVID-19 infections and mortality in Kenya, there has been a scarcity of resources like hospital facilities, quarantine centres and personal protective equipment (PPE) for the medical personnel. Therefore, effective planning was required by the Kenyan Government to ensure resources are available to combat the rising COVID-19 cases. This study developed the Autoregressive Integrated Moving Average (ARIMA) and the Holt-Winters models to predict the COVID-19 infections and mortality rates in Kenya. The Quantitative discrete data from the Kenya Ministry of health was used. The data covered 8 months' period from 23rd August 2020 to 23rd April 2021. The analysis entailed descriptive statistics and the time series prediction technique (ARIMA and Holt-Winters). The COVID-19 infections and mortality data were subjected to the time series prediction models to obtain the accuracy measures for model comparison. The study employed information criteria for model selection. The Simple linear regression was conducted on the COVID-19 data to determine a linear relationship between the COVID-19 infections recorded daily and the samples tested. The regression coefficients were obtained and they were statistically significant at 5% significance level. The R^2 of the model was 0.72 implying that 72% of the variation in the response variable is explained by the explanatory variable. The data were fitted to ARIMA and Holt-Winters models using R statistical software (version 4.1.0). The Final model with the lowest Akaike Information Criterion (AIC) was selected. The root mean square error was applied for model comparison and the ARIMA$(2,1,3)$ and the ARIMA$(0,1,1)$ were found to be the best models for predicting COVID-19 infections and mortality with minimum errors. The COVID-19 prediction was done at a 95% significance level using the selected ARIMA models. From the prediction plots, the infections and mortality cases were observed to increase significantly. Therefore, the study suggested that the ARIMA was an effective prediction model for the COVID-19 infections and mortality than the Holt-Winters model. The study recommended that Kenyans should observe the World Health Organization's guidelines to help reduce the infectivity and mortality rate. The government should provide protective equipment for the medical personnel to curb the surging COVID-19 cases. The study also recommended that the ARIMA model be applied for short-term prediction for further studies.

CHAPTER ONE

INTRODUCTION

1.1 Background Information

The COVID-19 (Coronavirus disease) is a respiratory disease and it originated from Wuhan in Dec. 2019 (Takele, 2020). The virus has now spread to the entire world and was declared a threat to human beings by the World Health Organization (WHO) due to its contagious nature. Aged people and people living with pre-existing medical conditions such as hypertension, diabetes and obesity are at a high risk of contracting the pandemic due to their weak immune systems (Wang et al., 2020). According to the WHO (2020), the COVID-19 is believed to have been caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2). According to (Maleki, et al., 2020), COVID-19 is transmitted from infected to susceptible individuals through respiratory droplets when in close contact. It also spread when people come into close contact with contaminated surfaces or objects. The virus can stay on surfaces for up to 12 hours before dying (Pandey et al., 2019). Maleki et al. (2020) suggested that the incubation period of for the COVID-19 is estimated to 14 days on human beings. The pandemic has affected every county's sector, including; the economy, transport, and education. According to the WHO report, there were over 454 million infected people, over 4 million fatalities globally as of 5th July 2021. Since 13th March 2020 when the first case of the COVID-19 infection was identified in Kenya, the infection cases have been increasing ever since.

The WHO (2020) recommended various measures to curb the surging COVID-19 infections and mortality. These measures include; regularly washing hands with soap and water, hand sanitizing, properly wearing a facemask, and keeping a social distance. This study developed the Autoregressive Integrated Moving Average (ARIMA) and the Holt-Winters models to predict the COVID-19 infections and mortality in Kenya. This study therefore compares the prediction accuracy of the ARIMA and the Holt-Winters models to determine an effective model for predicting the infection and mortality rates in Kenya.

The ARIMA and Holt-Winters are time series, prediction models. According to (Box et al., 2013), the ARIMA model was designed by Box and Jenkins to model stationary series.

Stationarity is the state in which both the mean and the variance of the series are constant over time. According to (Mathenge, 2019), the non-stationary series are made stationary through a differencing process of an appropriate order. The Augmented dickey fuller test was applied to determine the stationarity of the series. The Holt-Winters model was developed by peter winters to predict series with outliers thereby reducing the effects of random oscillations in time series (Habibur et al, 2016; Gelper et al, 2010). There are many models used to predict the transmission dynamics of infectious disease. Some of these modes include compartmental models such as Susceptible, Infected, and Recovered (SIR) and the Susceptible, Exposed, Infectious, and Recovered (SEIR), which involve dividing the entire population into departments, and agent-based models etc. The study used the ARIMA and the Holt-Winters models because they are the best time series models to predict the daily occurrences without the seasonality component since the COVID-19 quantitative discrete data was based on daily occurrences.

1.2 Problem Statement

There has been a rapid increase in the COVID-19 infections globally since the first case was reported. These increases in figures require proper and accurate prediction so that different countries Kenya being included can plan and manage the future spread of the pandemic. The ARIMA and Holt-Winters models were provided to give predictions with minimum errors hence providing accurate results. Although, no study had been conducted in Kenya to predict the COVID-19 infections and mortality using the Holt-Winters and ARIMA models simultaneously, this study, therefore, sought to model the COVID-19 infections and mortality rates in Kenya and access the prediction efficiency of the Holt-Winters model and the ARIMA model to identify an effective model for the COVID-19 prediction.

1.3 Justification of the study

The economic impacts, cost of treatment and mortalities caused by COVID-19 infections has led to scarcity of resources like hospital facilities, quarantine centres and personal protective equipment (PPE) for the medical personnel. Therefore, accurate planning had to be made by the government of Kenya to ensure that resources are made available to combat the rising COVID-19 cases. To ensure effective future planning for Combating the COVID-19 infections and mortality, a transmission model was developed for contingency planning and to inform the targeted prevention and disease control interventions.

1.4 Research Hypotheses

i. There is no linear relationship between the number of infections and the daily samples tested.

ii. There is no difference in the prediction accuracy between the ARIMA and Holt-Winters.

1.5 Significance of the study

This study was conducted to determine an effective model that can predict the COVID-19 infections and mortality in Kenya with minimum errors hence the government can adopt the prediction to make future planning concerning the spread of the pandemic and strategize measures of combating the pandemic. Also, the statisticians and the non-statisticians can be informed on the choice of a better forecasting model.

1.6 Objectives of the study

1.6.1 General Objective

The overall objective of this study was to model the COVID-19 pandemic in Kenya and access the prediction efficiency of the Holt-winters and the ARIMA models.

1.6.2 Specific Objectives

This study aimed to:

1. To fit a simple linear regression model on number of infected cases and samples tested.
2. To develop ARIMA and Holt-Winters models and use these model frameworks to predict COVID-19 infections and mortality in Kenya.
3. To evaluate the effectiveness of the ARIMA and the Holt-Winters models in predicting COVID-19 infections and mortality cases in Kenya.

1.7 Scope of the study

This study was conducted in Kenya using the COVID-19 quantitative discrete data ranging from 23rd Aug. 2020 to 23rd April 2021. The secondary data was collected by the ministry of health. The study was aimed at accessing the forecasting accuracy of the models. The data used comprised of the daily infected cases, daily death cases and the daily samples tested.

CHAPTER TWO

LITERATURE REVIEW

2.1 Modelling the COVID-19 pandemic

Different studies have been done worldwide by different researchers to model and predict the spread of the COVID-19 pandemic by applying different existing models. Zhou et al. (2020) applied the Susceptible Exposed Infectious and Recovered (SEIR) and logistic stochastic models to predict the worldwide COVID-19 pandemic. After performing several differentials equations, the following logistic regression model was formulated.

$$p(t) = \frac{kp_0 e^{rt}}{k + p_0\left(e^{rt} - 1\right)}$$

Where, r represent the growth rate, k represent the case volume, $p(t)$ represent the final value, and p_0 represent the initial value. The SEIR and the logistic regression models were compared and it was discovered that the SEIR model performed better than the logistic regression model. It was also found that reducing the contact rate minimizes the spread of the virus.

Zhao et al. (2020) applied the maximum hasting (MH) parameter estimation and modified SEIR to model and predict the COVID-19 epidemic in African countries. The findings revealed that total quarantine, control of movement, expansion of the healthcare system, and strict government enforcement of anti-epidemic measures might prevent the pandemic’s spread.

Barbastefano et al. (2020) discussed the ordinary differential equation (ODE) and applied to predict the COVID-19 pandemic among the top affected countries of the world. The formulated differential equation was given by,

$$N_0 = S(k) + Q(k) + E(k) + A(k) + I(k) + M(k) + D(k) + R(k)$$

Where;

$$M(k) = N_0 - S(k) + Q(k) + E(k) + A(k) + I(k) + D(k) + R(k)$$

The parameters were estimated by applying the nonlinear least-squares minimization and curve-fitting methods. From the findings, it was concluded that the developed ordinary differential equation predicted the COVID-19 pandemic perfectly well and it was suggested that people across the world should adopt WHO guidelines to help reduce the infection rate.

Maleki et al. (2020) applied the time-series model to predict the COVID-19 registered and recovered cases in the affected countries of the world. The developed Autoregressive prediction model was given by,

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + z_t$$

From the findings, it was concluded that the developed model predicted the confirmed and recovered cases with minimum errors. It was therefore concluded that the proposed model was more accurate than the standard Gaussian model based on model selection criteria.

Reza et al. (2020) applied the time series analysis technique to model and predict the spread of the COVID-19 and the mortality rate in the world. The applied ARMA prediction model was given by,

$$x_t = \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + z_t + \eta_1 z_{t-1} + \cdots + \eta_q z_{t-q}$$

Where, $z_t \sim WN(0, \sigma^2)$

The maximum likelihood estimator was applied to estimate the model parameters.

$$L(\theta) = f_x(x \mid x_0, z_0, \theta) = \prod_{t=1}^{n} g(z_t \mid x_0, z_0, \theta)$$

Where, $L(\theta)$ denotes the likelihood conditional function. It was concluded that the ARMA model predicted the spread of COVID-19 perfectly, and therefore, the study suggested that the model be applied for forecasting in further studies.

Singh et al. (2020) discussed the comparison of the Hybrid-wavelength ARIMA model and the econometric ARIMA model to forecast the COVID-19 pandemic. The developed ARIMA(p, d, q) model was defined by,

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^d = \left(1+\sum_{j=1}^{q}\theta_j L^j\right)\varepsilon_j$$

The two models were compared and it was discovered that the Hybrid-ARIMA was more accurate for the prediction of the COVID-19 pandemic than the econometric ARIMA model. Therefore, the study suggested that the Hybrid-ARIMA model be used for forecasting the pandemic because, according to the results, the Hybrid ARIMA model was 80% effective than the econometric ARIMA model.

Khrapov & Loginova (2020) developed and applied the SIR mathematical model to predict the dynamic of Coronavirus epidemic development in China. The SIR epidemiological model was expressed by the following equation,

$$N = S(t) + I(t) + R(t)$$

Where, $S(t)$ is the healthy individuals who are at risk of contracting the virus, $I(t)$ is the infected individuals and $R(t)$ represent the individuals who have died and those who have recovered. The following differential equations were developed to help in the modelling of the virus.

$$\frac{dS(t)}{dt} = -\alpha S(t) + I(t)$$

$$\frac{dI(t)}{dt} = \alpha S(t) I(t) - \beta I(t)$$

$$\frac{dR(t)}{dt} = \beta I(t)$$

In the above, α and β are the model parameters. Based on the findings, it was concluded that the epidemic would peak in February 2021.

Benvenuto et al. (2020) applied the ARIMA model to predict the COVID-19 epidemic trend in the world. The moving average and autoregressive average order were estimated by applying the ACF plot and the PACF plot. The formulated ARIMA model was given by,

$$x_t = \gamma + \beta_1 x_{t-1} + \cdots + \beta_p x_{t-p} + \phi_t + \theta_1 \phi_{t-1} + \cdots + \theta_q \phi_{t-q}$$

According to the results, the ARIMA$(1,0,3)$ was identified as the best prediction model for the incidences of the COVID-19 pandemic since it had minimum AIC. Also, from ACF and PACF, it was discovered that the prevalence and incidence of the COVID-19 epidemic were not influenced by the seasonality component. The study concluded that there was a possibility that the curve will flatten around December 2020.

Wang et al (2020) applied the SEIR model to estimate the number of COVID-19 infected cases among the Chinese. The differential equations of the SEIR model were given by the following equations,

$$\frac{dS}{dt} = \lambda S \big/ N$$

$$\frac{dE}{dt} = \lambda S \big/ N - \delta E$$

$$\frac{dI}{dT} = \delta E - \xi I$$

$$\lambda = R_0 \xi$$

In the above, λ denotes the transmission rate δ denotes the infectivity rate and ξ denotes the recovery rate. The findings of the study revealed that the transmission rates will continue to rise and therefore, it was suggested that the Chinese should adopt the WHO guidelines to curb the spread of the pandemic.

Rajagopal et al. (2020) applied the fractional symptomatic infectious recovered and deceased (SEIRD) ordered model to predict the novel COVID-19 outbreak. The fractional SEIRD ordered model was expressed by the following equation.

$$D^q f(t) = \frac{1}{\Delta(n-q)} \int_0^t (t-\tau)^{n-q-1} f^n(\tau) d\tau = j^{(n-q)} \left(\frac{d^n}{dt^n} f(t) \right)$$

The parameters were estimated by applying the RMSE and it was computed mathematically as;

$$RMSE = \sqrt{\frac{\sum_{i=1}^{T} (I_r - I_n)^2 T (D_r - D_m)^2}{T}}$$

The study concluded that the fractional SEIRD ordered model provided a better result than the integer-ordered model due to minimum errors. It was suggested that the fractional SEIRD ordered model be used for daily forecasting.

Xiao & Wu. (2020) applied the ARIMA model to predict the spread of the COVID-19 pandemic among the most affected countries of the world. The Box-Jenkins approach was developed by combining the Autoregressive process with the moving average process.

$$AR(P): Z_t = C + \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + \epsilon_t$$

$$MA(q): Y_t = \mu + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

The developed ARMA model was represented by the below equation,

$$Z_t = C + \mu + \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

The method of Least Square Estimator (LSE) was applied to estimate the model parameters. The study applied the mean absolute deviation and the mean absolute percentage error to estimate the model accuracy. The study suggested that the ARIMA model provided better prediction for the COVID-19.

Yang et al. (2020) applied the ARIMA model to predict the transmission of the COVID-19 epidemic in Italy. The ARIMA prediction model was given by,

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j e_{t-j}$$

In the above, p represents the AR order, q is the MA order and X_t is the model parameter. The $\text{ARIMA}(2,2,1)$ was identified a better prediction model for the COVID-19 transmission with an R^2 value of 0.956. The study concluded that the model fitted well and was considered suitable for short term prediction.

Mohammed & Gupta (2020) Compared the prediction accuracy of the ARIMA and the Nonlinear Autoregressive Neural Network (NANN) for the COVID-19 pandemic. The $\text{ARIMA}(1,1,0)$ was identified the best prediction model for the COVID-19 pandemic based on the information criterial. The R^2 for the ARIMA and the NANN models were given by 0.95 and 0.97 respectively. The study concluded that the Nonlinear Autoregressive Neural Network model yielded accurate results for predicting the COVID-19 pandemic than the ARIMA model.

Zhao et al. (2020) discussed the application of Time series analysis to model the novel Coronavirus in Wuhan China. The study applied the Holt-Winters smoothing method and the ARIMA model. The following equations were formulated,

$$S_k = \beta Y_k (1-\beta)(S_{k-1} + T_{k-1})$$

$$T_k = \alpha(S_k - S_{k-1}) + (1-\alpha)T_{k-1}$$

$$\hat{Y}_{k(m)} = S_k + mT_k$$

k represents the current period, Y_k is the actual observation in period k, T_k is the forecasted trend at time k, S_k is the estimated level at period k, $\hat{Y}_{k(m)}$ is the previously estimated value, β and α are smoothing parameters and m is the number of the forecast. Based on the findings, it was suggested that the Holt-Winters and the ARIMA models successfully predicted the pandemic.

Bhattacharya & Banerjee (2020) applied the exponential growth model and the Box-Jenkins approach to estimate and predict the COVID-19 pandemic incidences in India. The formulated models were expressed as,

$$N(k) = N(0)e^{\eta k}$$

$$Z_t = \eta + \sum_{i=1}^{p} \vartheta_i \upsilon_{t-i} + \sum_{i=1}^{q} \phi_i \upsilon_{t-i} + \upsilon_t$$

Where, η is the mean, ϑ and ϕ are the estimated parameters. The stationarity of the series was determined by applying the Augmented dickey fuller's test. For model selection, AIC was applied to select the appropriate ARIMA model, in which case the ARIMA$(5,2,1)$ was Identified as the best model. The study concluded that the ARIMA$(5,2,1)$ forecasted daily confirmed cases appropriately and suggested that the government and health care should use the predicted results to lay down guidelines that will help curb the further spread of the pandemic.

Tinani et al. (2020) applied the Autoregressive Integrated Moving Average (ARIMA) model to analyzed and forecasted the COVID-19 pandemic among India's hotspot states. They developed three ARIMA models, i.e., for confirmed cases, recoveries and fatalities respectively.

Death model: $Y_t = \psi + \vartheta_1 y_{t-1} + \cdots + \vartheta_p y_{t-p} + \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \cdots + \Theta_q \varepsilon_{t-q}$

Y_t represents the forecasted fatalities, ψ is the model intercept, ϑ and Θ are estimated parameters.

Recovery model: $Z_t = \zeta + \gamma_1 z_{t-1} + \cdots + \gamma_p z_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \cdots + \phi_q \epsilon_{t-q}$

Z_t represents the forecasted recoveries, ζ is the mean of the model, γ and ϕ are the estimated parameters and finally, the ARIMA model for the confirmed positive cases is;

$Z_t = \Lambda + \alpha_1 z_{t-1} + \cdots + \alpha_p z_{t-p} + \varepsilon_t + \Phi_1 \varepsilon_{t-1} + \cdots + \Phi_q \varepsilon_{t-q}$

Where, z_t are the forecasted positive confirmed cases, Λ is the model intercept, Φ and α are estimated model parameters. The AIC technique was applied for model selection and the ARIMA$(2,2,0)$, the ARIMA $(1,2,2)$ and the ARIMA$(5,2,0)$ for fatalities, recoveries and confirmed cases respectively were identified as the best models with minimum AIC. Based on the findings, they suggested that the Government should adopt their forecast to curb the further spread of the virus.

According to (Chintalapudi et al., 2020) ARIMA model was applied to forecasted COVID-19 registered and recovered cases in Italy. They formulated the non-seasonal ARIMA model due to the absence of seasonality in the series. The formulated model was given by;

$$Y_t = c + \phi_1 yd_{t-1} + \cdots + \phi_p yd_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$$

The best selected ARIMA models were the ARIMA$(1,2,0)$ and the ARIMA$(3,2,0)$ for infectivity and recovery with accuracy of 93.75% and 84.4% respectively. The study suggested that the Health ministry and the national government should put strict measures to mitigate the pandemic spread.

Kumar et al. (2020) forecasted the COVID-19 pandemic in India. They applied different forecasting models. ARIMA, SEIR, and Richard's models. The formulated ARIMA model was defined by,

$$\acute{y}_t = C + \phi_1 \acute{y}_{t-1} + \cdots + \phi_p \acute{y}_{t-p} + \theta_1 y_{t-1} + \cdots + \theta_q y_{t-q} + \varepsilon_t$$

In the above, $y_t^{'}$ is the differenced series, and the best-selected models were found to be ARIMA$(1,2,0)$ and ARIMA$(0,2,2)$ for modelling the confirmed positive cases and fatalities respectively. Richards's model was given by,

$$W(t) = \frac{\alpha}{\left(1+\beta exp(-kt)\right)^{1/m}} + \varepsilon$$

Where, β is the growth rate, k represents the observations made on the data, t represents the time taken to observe, m denotes the slope whereas, α denotes the upper value The SEIR model was defined by four different equations.

$$dS/dt = -\lambda IS$$

$$dI/dt = \lambda IS - \gamma I$$

$$dR/dt = \gamma I$$

$$\lambda = CP$$

S is the individuals susceptible to the pandemic, I is the individuals infected with the virus, R is the individuals that have recovered and died, λ is the infectivity rate, C is the contact rate and P is the probability of transmission and γ is the recovery rate. They used R^2, RMSE, and BIC for model validation, and they discovered that ARIMA provides a better prediction estimate for the COVID-19 pandemic compared to Richard's and SIR model

Gathungu et. al. (2020) applied the SEIR to model to the effect of the non-pharmaceutical interventions on the transmission of the COVID-19 epidemic in Kenya. The mathematical model developed was expressed as;

$$N(k) = S(k) + E(k) + A(k) + M(k) + H(k) + C(k) + D(k) + R(k)$$

In the above model, $S(k)$ represent susceptible individuals, $E(k)$ represent exposed individuals, $M(k)$ represent symptomatic individuals, $A(k)$ represent asymptomatic infectious individuals, $H(k)$ represent severe and hospitalized individuals $C(k)$ represent critically ill individuals, $R(k)$ represent individuals that have recovered from the pandemic, $D(k)$ represent fatalities and $N(k)$ represent the total human population is likely to be infected with the disease. Based on their findings, they concluded that closing schools, dusk to dawn curfew, and partial lockdown of affected areas would help prevent the COVID-19 pandemic from spreading.

Odhiambo et al., (2020) applied the compound Poisson regression model to predict The COVID-19 spread in Kenya. The regression Poisson model developed was given by,

$$Z_k \sim pois\left(\sum_{k=1}^{\infty}\varphi_k \vartheta\right)$$

Where, $\varphi_k \vartheta$ is the mean parameter of the registered COVID-19 cases. The method of ordinary least square (OLS) estimator was applied to estimate the model parameters. It was observed that the Poisson regression model predicted the COVID-19 pandemic in Kenya perfectly well. The study concluded that the Government should put strict measures like increasing testing centres to ensure they curb the pandemic's further spread.

2.2 Model Comparison

Over the years, different studies have been done to determine a better forecasting model between the Holt-Winters and the ARIMA. Rahman & Ahmar (2016) forecasted energy consumption in the USA and compared the accuracy measures of the Holt-Winters and the ARIMA models to determine a better prediction model. Based on the MAE and the MASE, the study concluded that Holt-Winters was more effective for predicting the energy consumption compared to the ARIMA model due to minimum errors. Mini et al. (2015) discussed modelling CPU for fishery along the coast of India. They compared ARIMA, Holt-Winters and NNAR models. From the study, their findings revealed that the ARIMA model provided a better fit for the prediction of CPU series based on the MAPE.

Barr et al. (2021) applied the ARIMA, the Poisson and the Holt-Winters smoothing models to predict the confirmed and death COVID-19 cases. The MSE was used for the comparison and the ARIMA model was found to be a better model for prediction with a minimum MSE of 0.079. Omane-adjepong et al. (2013), compared Seasonal ARIMA and Holt-Winters to determine an effective forecasting model for the Ghana’s inflation rate. The forecasting accuracy was measured using MAE, MAPE and MSE. Based on the results, they concluded that Seasonal ARIMA provided a better estimate for forecasting Ghana inflation. Makatjane & Maroke (2016) compared Holt-Winters and ARIMA model for forecasting car sales in South Africa and their findings revealed that Holt-Winters provided a better prediction estimate compared to the ARIMA model.

The Holt-Winters and the Box-Jenkins models were compared to predict software failures (Yakovyna & Bachka, 2018). Based on their findings, the MAPE revealed that the ARIMA model provided a better prediction estimate for software failures than Holt-Winters. Veiga et. al (2014) forecasted food demand in retail and made a compared the prediction efficiency of the Holt-Winters and the ARIMA models. The prediction measures were obtained and based on the RMSE, the Holt-Winters model was found to provide a better forecasting performance and hence preferred for forecasting food demand. The Seasonal Autoregressive Integrated Moving Average (SARIMA) model and the Holt-Winters model were applied to forecast the Inflation rate in Kenya (Lidiema, 2017). The Mean Absolute Error and the Mean Percentage Error were applied for model comparison. The findings of the study suggested that the Seasonal ARIMA model was more accurate and therefore, a better prediction model.

2.3 Research gap

Since the COVID-19 pandemic emerged, several studies have been conducted to determine the most effective model for predicting the COVID-19 transmission. According to studies done by Habibur et al. (2016), Mini et al. (2015), Omane-adjepong.et al. (2013), (Makatjane & Moroke, 2016), (Veiga et al., 2014), (Yakovyna & Bachka, 2018), Barr et al. (2021) and (Lidiema, 2017), the Holt-Winters and the ARIMA models were compared to determine an prediction model for COVID-19 variant. The accuracy measures such as the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE) and the Mean Absolute Square Error (MASE) were applied for model comparison. The findings of these studies gave inconsistent results thereby finding it difficult to determine an effective prediction model based on the MAE, MAPE and MASE. Therefore, this study proposed to apply the RMSE as the accuracy measure for comparing the forecasting efficiency between the ARIMA and the Holt-Winters models and determine the appropriate model for the prediction of the COVID-19 transmission and mortality in Kenya. The RMSE was applied for model comparison because, according to (Makridakis, 1993; Mahmoud, 1984; Davydenko & Fildes, 2013), they suggested that the RMSE was a better accuracy measure for model comparison with minimum errors.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Data Description

The COVID-19 quantitative discrete data was a secondary data collected by the health officials and it was retrieved from the ministry of health official website, https://www.health.go.ke The COVID-19 data was available for daily infection cases, fatalities and samples tested from 23rd Aug. 2020 to 23rd July 2021.

3.2 Data Analysis

A simple linear regression was done on the COVID-19 data to determine the linear relationship between the daily samples tested and the daily number of infected cases. The COVID-19 infections and mortality data were fitted to the Holt-Winters and the ARIMA models with Root Mean Square Error (RMSE) being used to select the final model. All the models were fitted using R statistical software version 4.1.0. The following packages were used while fitting the data, ggplot2, forecast, astsa, readxl, lmtest and tseries. The following functions were also used while modelling the COVID-19 infections and mortality ndiffs (), auto. arima (), HoltWinters (), Box.test () and adf.test ().

3.3 Simple Linear Regression

The Simple linear regression determines a linear relationship between the response variable and the explanatory variable. The simple linear regression model between COVID-19 infections and daily samples tested was given by,

$$Y = \beta_0 + \beta_1 X \tag{3.1}$$

Where, Y is the response variable and X is the explanatory variable.

3.4 ARIMA Model Description

3.4.1 Autoregressive Process

A stochastic process $\{z_t, t \geq 0\}$ is an autoregressive process of order p, $\mathrm{AR}(p)$ defined by;

$$Z_t = \vartheta_1 z_{t-1} + \vartheta_2 z_{t-2} + \cdots + \vartheta_p z_{t-p} + \varepsilon_t \tag{3.2}$$

Where; p is an autoregressive order, $\vartheta_1, \cdots, \vartheta_p$ are the estimated parameters of autoregressive terms and ε_t is a series of white noise with mean 0 and variance σ^2. By applying the backshift operator, the autoregressive equation is expressed as,

$$z_t = \vartheta(B)t \tag{3.3}$$

Where, $\vartheta(B) = 1 - \vartheta_1 B - \vartheta_2 B^2 - \cdots - \vartheta_P B^P$

3.4.2 Moving Average process

A Moving average process z_t denoted as $\mathrm{MA}(q)$ is defined by;

$$Z_t = \varphi_1 e_t + \varphi_2 e_{t-1} + \varphi_3 e_{t-2} + \cdots + \varphi_q e_{t-q} \tag{3.4}$$

From the equation above, q is the moving average order, $\varphi_1, \cdots, \varphi_q$ are the parameters of the MA process and e_t represents the error term with a mean of 0 and a variance of σ^2. The backshift operator for the moving average of order q was expressed by,

$$\varphi(B) = 1 + \varphi_1 B + \varphi_2 B^2 + \cdots + \varphi_P B^P \tag{3.5}$$

3.4.3 Autoregressive Moving Average process

ARMA process was formed by combining the $\mathrm{AR}(p)$ model and the $\mathrm{MA}(q)$ model. The $\mathrm{ARMA}(p,q)$ process was defined by the below equation 3.5;

$$Z_t = \vartheta_1 z_{t-1} + \vartheta_2 z_{t-2} + \cdots + \vartheta_p z_{t-p} + \varphi_2 e_{t-1} + \varphi_3 e_{t-2} + \cdots + \varphi_q e_{t-q} + e_t \tag{3.6}$$

The backshift operator for the ARMA process is given by; $\vartheta(B) z_t = \varphi(B) e_t$

From $\mathrm{MA}(q)$ and $\mathrm{AR}(p)$, the backshift operator for the ARMA process was simplified to,

$$\left(1 - \sum_{i=1}^{P} \vartheta_i B^i\right) z_t = \left(1 + \sum_{j=1}^{q} \varphi_j B^j\right) e_t \tag{3.7}$$

3.4.4 Autoregressive Integrated Moving Average process

The ARIMA model is an extension of the ARMA model to a non-stationary time series. The ARIMA model of order p, d and q is expressed as ARIMA(p,d,q) where p denotes the Autoregressive order, d is the differenced or integrated part and q denotes the Moving Average order. The ARIMA model is widely applied for prediction because it can handle any series with or without the seasonality component (Mathenge, 2019). When the ARIMA model is differenced d times, the process becomes an ARMA process. For instance, if z_t is non-stationary, then the first difference will be $\Delta z_t = z_t - z_{t-1}$ So that the differenced ARMA model will be given by;

$$\Delta z_t = \vartheta_1 \Delta z_{t-1} + \cdots + \vartheta_p \Delta z_{t-p} + \varphi_1 e_{t-1} + \cdots + \varphi_q e_{t-q} + e_t \tag{3.8}$$

3.5 Holt-Winters modelling

The Holt-Winters prediction model was developed to predict series with outliers. It uses weighted historical trend to predict future values and is made up of two components i.e., the smoothing constant (E,α) and the trend component (T,β). The Holt-Winters forecasting model is obtained as,

$$\xi_{t+k} = E_t + \kappa T_t \tag{3.9}$$

Where E is the intercept and T is the slope and they are computed recursively as follows;

$$E_t = \alpha(x_{t-1}) + (1-\alpha)(E_t + T_{t-1})$$
$$T_t = \beta(E_t - E_{t-1})(1-\beta)T_{t-1}$$

Where; ξ_{t+k} is the forecasted value of k from period t, x_{t-1} is the real value, E_{t-1} is the estimated value, T_{t-1} is the trend value, α is the smoothing constants for estimates, β is the smoothing constant for the trend and κ is the period. The coefficients α and β lies between 0 and 1 i.e. $0 \le \alpha \le \beta \le 1$

3.6 Box-Jenkins Framework

Box et al. (2013) proposed 4 iterative stages which include; model Identification, Estimation, diagnostic test and forecasting as shown in figure 1.

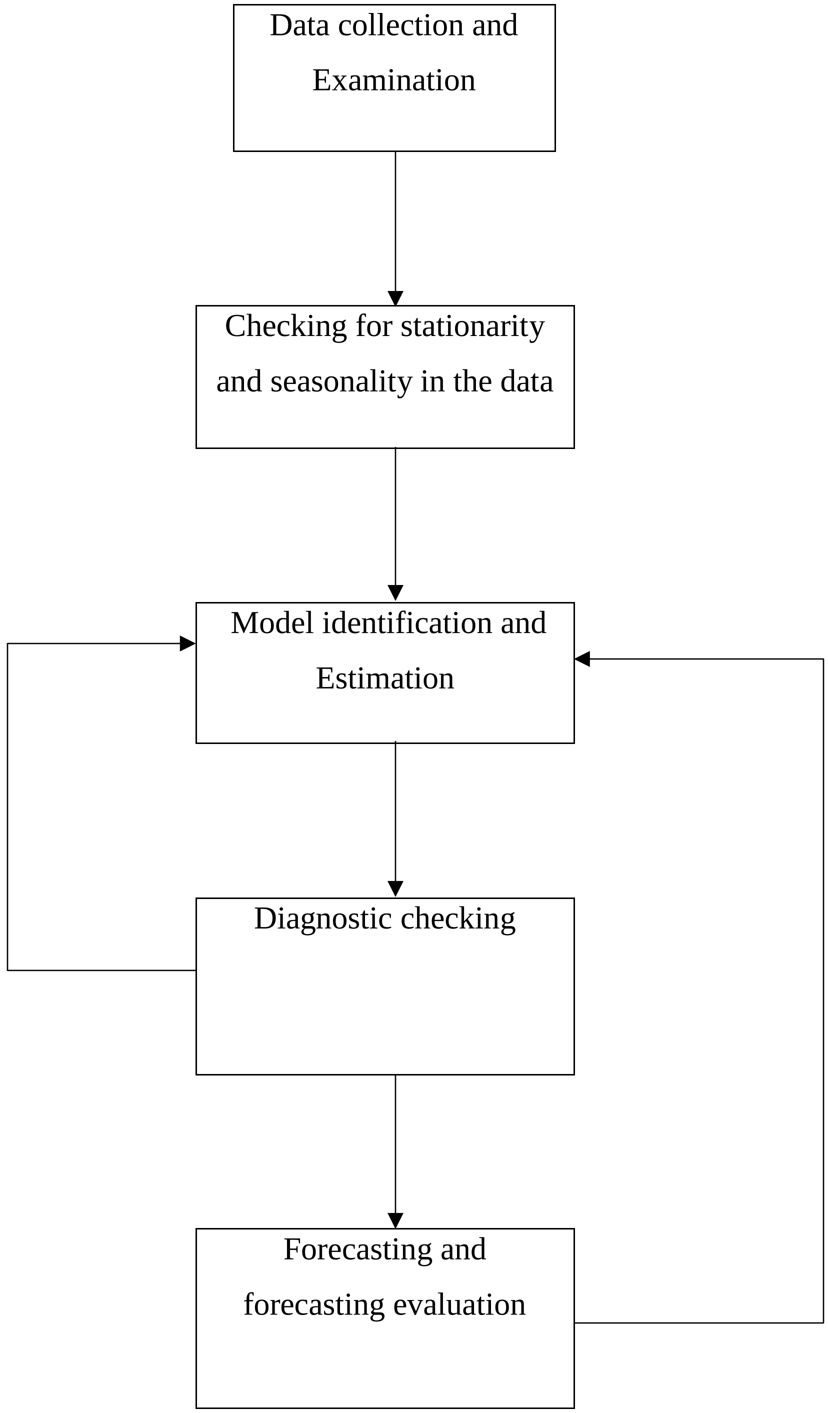

Figure 1: ARIMA forecasting procedure

3.6.1 Model Identification

This is done by checking for stationarity or seasonality in the series and choosing a specified model. Firstly, the series is assumed to be non-stationary, then the ACF and the PACF were plotted to confirm the assumption. The ACF plot with a slow decay indicates non-stationarity in the series. According to this research, stationarity was checked using the Augmented Dickey Fuller's (ADF) test.

3.6.2 Model Estimation

The coefficients that best fit the selected ARIMA model are estimated consistently by the maximum likelihood estimate (MLE) or least-square estimate (LSE). This study employed the MLE technique to estimate the parameters of the model. The mathematical derivation of MLE was computed as follows,

Consider the model given below,

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} - \vartheta_1 \varepsilon_{t-1} - \cdots - \vartheta_q \varepsilon_{t-q} - \varepsilon_1 \tag{3.10}$$

The Likelihood function is the joint density function for all observations for given values of ϑ and, σ_{ε}^{2}.

The likelihood function of ϑ and σ_{ε}^{2} is given by $L\left(y_N : \vartheta, \sigma_{\varepsilon}^{2}\right) = f\left(y_N \mid \vartheta, \sigma_{\varepsilon}^{2}\right)$

Given the observation, y_N we estimate ϑ and σ_{ε}^{2} the values for which the likelihood is minimized. ε_N is a white noise process at time N and the variables $y_N \mid y_{N-1}$ and y_N are independent, hence,

$$f\left(y_N \mid \vartheta, \sigma_{\varepsilon}^{2}\right) = f\left(y_{N-1}, \vartheta, \sigma_{\varepsilon}^{2}\right) f\left(y_{N-1} \mid \vartheta, \sigma_{\varepsilon}^{2}\right)$$

Therefore;

$$L\left(y_N, \vartheta, \sigma_{\varepsilon}^{2}\right) = \left(\prod_{t=p+1}^{N} f\left(y_t \mid y_{t-1}, \vartheta, \sigma_{\varepsilon}^{2}\right)\right) f\left(y_p \mid \vartheta, \sigma_{\varepsilon}^{2}\right)$$

The process is assumed to be distributed as a Gaussian process then,

$$f\left(y_t \mid y_{t-1},\vartheta,\sigma_\varepsilon^{\ 2}\right)=\frac{1}{\sigma_\varepsilon\sqrt{2\pi}}exp\left(-\frac{1}{2\sigma_\varepsilon^{\ 2}}\left(y_t \mid y_{t-1}\left(\vartheta\right)\right)^2\right)$$

Therefore, $L\left(y_N,\vartheta,\sigma_\varepsilon^{\ 2}\right)=\left(\sigma_\varepsilon^{\ 2}2\pi\right)^{\frac{N-p}{2}}exp\left(-\frac{1}{2\sigma_\varepsilon^{\ 2}}\sum_{t=p+1}^{N}\varepsilon_t^{\ 2}\left(\vartheta\right)\right)$

The MLE $\hat{\vartheta}$ is a prediction error estimate since it is obtained by minimizing

$$S\left(\vartheta\right)=\left(\sum_{t=p+1}^{N}\varepsilon_t^{\ 2}\left(\vartheta\right)\right)$$

By differentiating $S\left(\vartheta\right)$ with respect to $\sigma_\varepsilon^{\ 2}$, it can be shown that the MLE of $\sigma_\varepsilon^{\ 2}$ is given by; $\hat{\sigma}_\varepsilon^{\ 2}=S\left(\hat{\vartheta}\right)\Big/\left(N-p\right)$. Therefore, the estimate $\hat{\vartheta}$ is asymptotically unbiased and efficient.

3.6.3 Model Selection

The Akaike Information Criterion was applied to select the best model from a set of different models in a given set of data. The AIC is best model selection criterion. According to (Maurice, 2019), the Akaike Information Criterion estimates the quality of a model in a set of data. The model with the lowest AIC value was selected as the best model because it minimizes the errors. The criterion is computed as;

$$AIC=-2logL+2k \tag{3.11}$$

In the above equation 3.12, T denotes the number of observations, L denotes the likelihood function whereas k denotes the estimated parameters of the model.

3.6.4 Forecasting Accuracy

The performance of a model was assessed by considering its forecasting accuracy. The best prediction model is the one with minimum errors. The forecasting accuracy between the ARIMA and the Holt-Winters model was compared using the Root Mean Square Error (RMSE) which was computed mathematically as;

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\hat{X}_t - X_t\right)^2} \tag{3.12}$$

3.6.5 Diagnostic Checking

The selected model was tested to determine whether it conforms to the specifications of stationarity. After fitting the model and estimating parameters, the diagnostic test was performed. According to (Keith, W., & Mcleod, A., 1994), Diagnostic testing in time series is applicable when parameters were estimated using the maximum likelihood estimate. According to this research, the diagnostic checking was done using residuals. The Q-Q plot and histogram for the residuals were designed to test for the normality of the model. The bell-shaped histogram implied that the selected model was normal. The Box Ljung test statistic was applied to test for the presence of serial correlation between the lags or determine if the residuals are independent and identically distributed. Ljung-Box test statistic was computed as,

$$Q^* = n(n+2)\sum_{k=1}^{K}\frac{\hat{\rho}_k^{\,2}}{n-k}. \tag{3.13}$$

From the equation 3.14, n is the observations and k is the lag. The test statistic Q^* is distributed as a chi-square.

3.6.6 Forecasting

Forecasting in time series is the technique of estimating future events based on past and present data. After performing a diagnostic test and the model has passed the test, forecasting was then performed.

CHAPTER FOUR

RESULTS

4.1. Introduction

The data analysis discussed in this chapter includes; the Regression analysis and the time series analysis (ARIMA and Holt-Winters).

4.2 Simple Linear Regression

A simple linear regression was conducted to examine the linear relationship between the number of infected cases and the daily samples tested. Table 1 below shows the simple linear regression coefficient.

Table 1: Regression Coefficient

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	-99.170	23.383	-4.241	4.38e-05
Sample size	0.109	0.006	17.639	< 2e-16

The regression coefficient between the number of infected cases and the daily samples tested was -99.170 and 0.109 respectively. The response variable was statistically significant at a 5% confidence level implying a linear relationship. therefore, the null hypothesis which stated that there was no linear relationship between the number of infected cases and the daily samples tested was rejected implying the existence of a linear relationship between the two variables. The R-Squared was 0.72 implying that 72% of the variation in the response variable is explained by the explanatory variable. The linear regression equation was given by,

$$Infected\ cases = -99.170 + 0.109 * Sample\ size \tag{3.14}$$

The scatterplot was also designed to visualize the linear relationship between the infected cases and the daily samples tested. From the plot in Figure 2, it was observed that there was a linear relationship between the number of infected cases and the daily samples tested. This implied that an increase in the sample size increases the number of infected cases recorded daily.

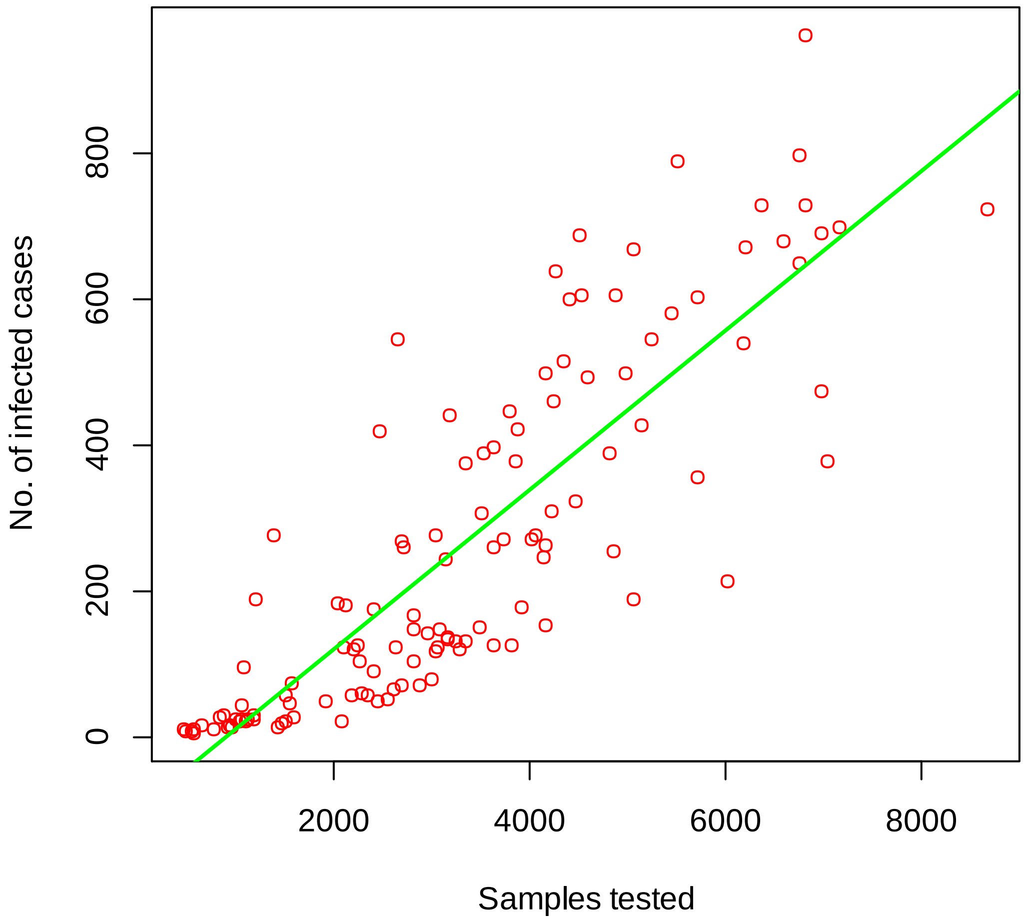

Figure 2: Scatterplot of infected cases vs samples tested

4.3 Trend Analysis

Figure 3 and 4 below shows the trend analysis of the COVID-19 mortality and infection in Kenya. From the plots, a linear trend was observed and the linear trend equations for the infected and mortality cases were given by, $y_t = 87.65 + 2.657x_t$ and $y_t = 1.463 + 0.04014x_t$ respectively.

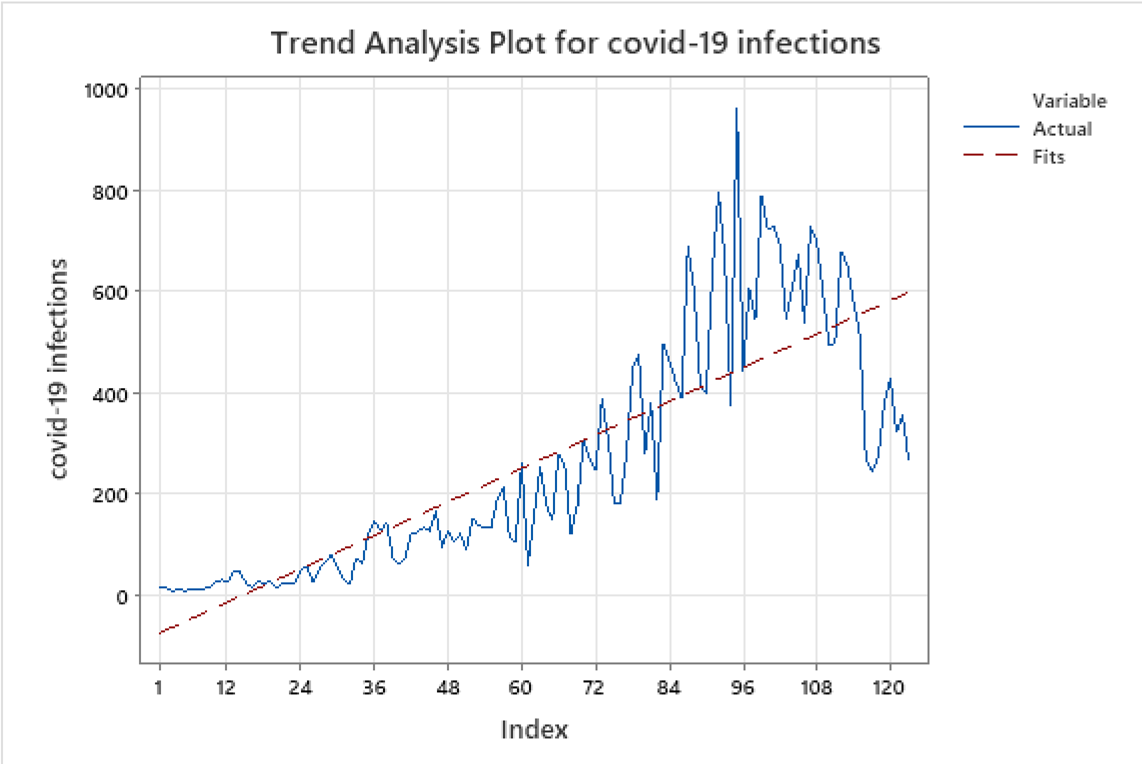

Figure 3: Plot of trend analysis for COVID-19 infections

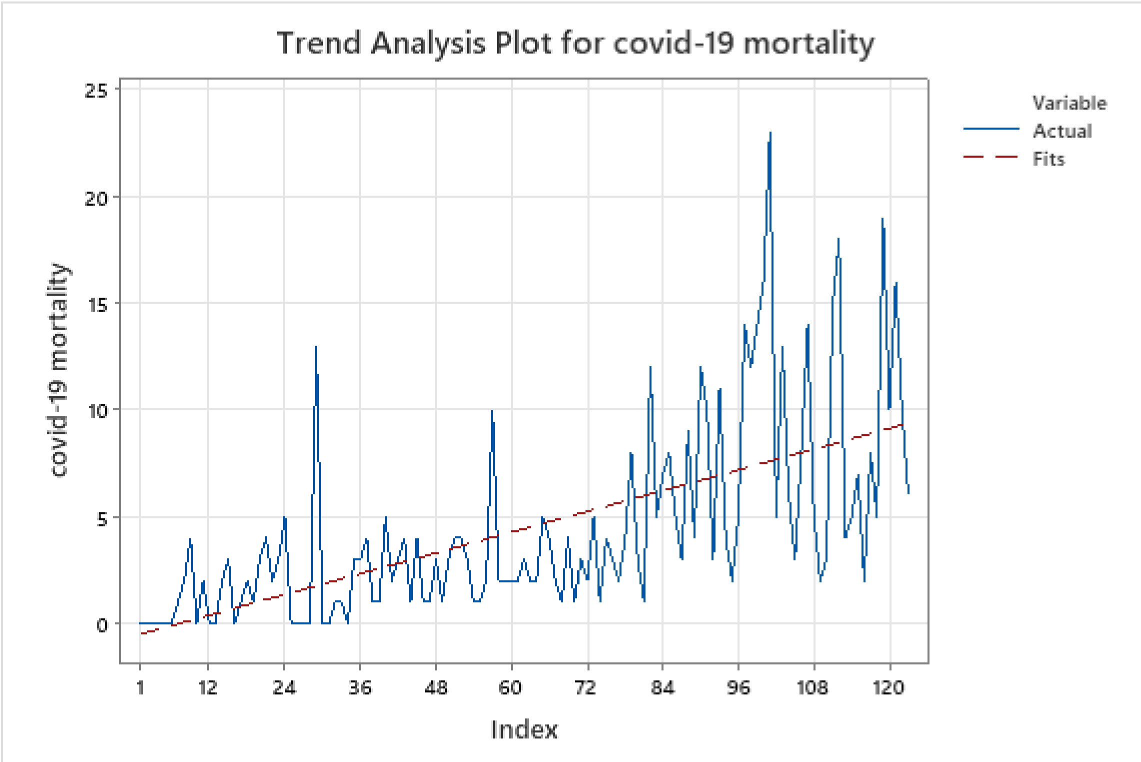

Figure 4: Plot of trend analysis for COVID-19 mortality

4.4 Model Identification

Figure 5 and 6 below shows the plot for the COVID-19 infections and mortality respectively. From the plots, it was observed that both the series were non-stationary due to the presence of a trend.

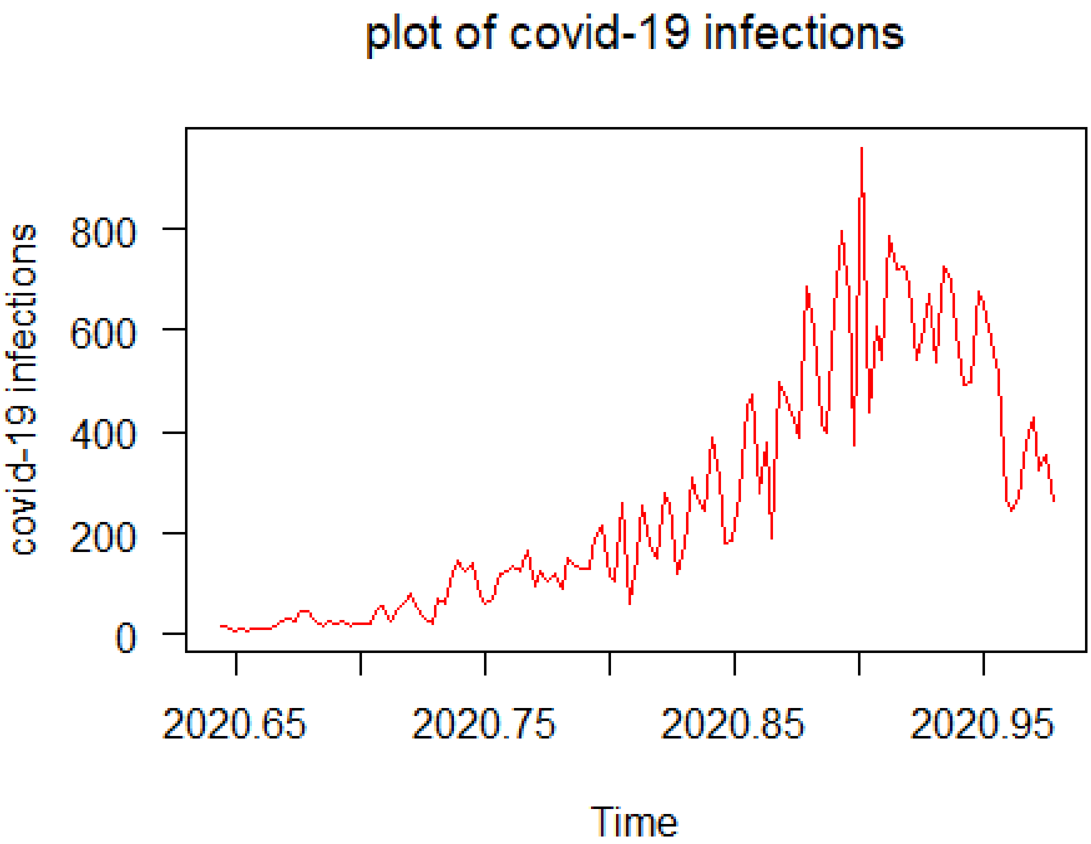

Figure 5: Time series plot for COVID-19 infections

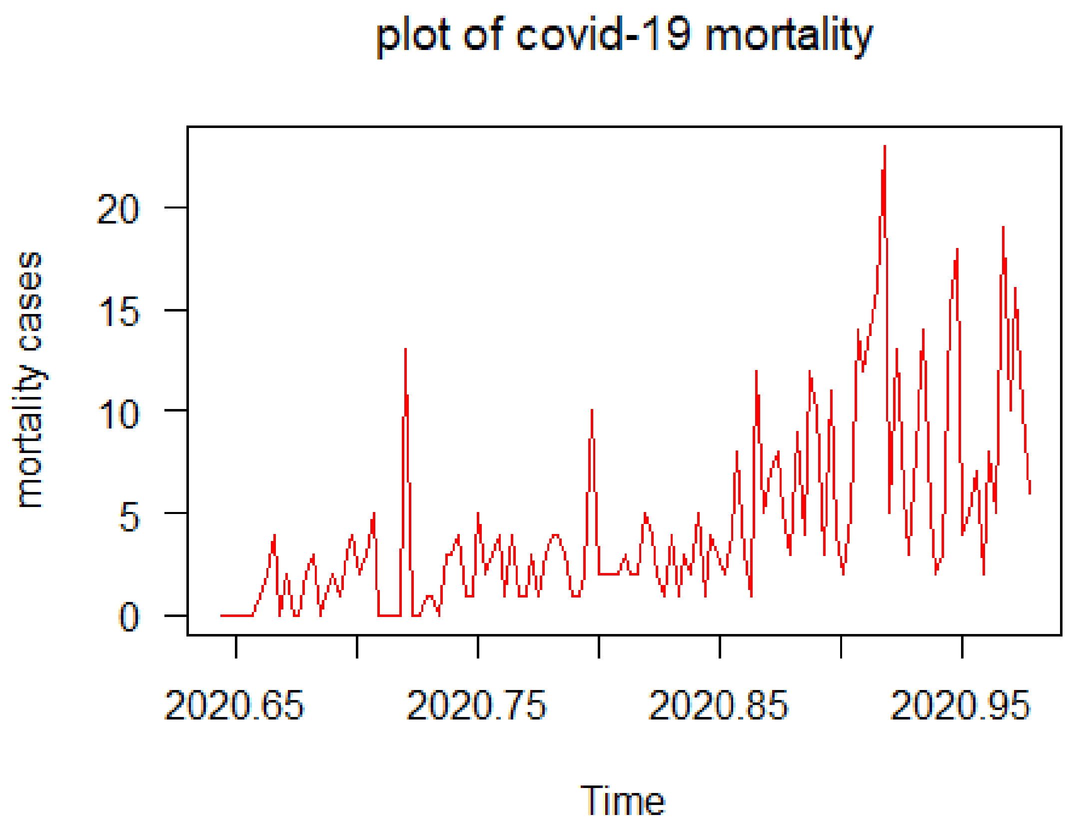

Figure 6: Time series plot for COVID-19 mortality

The Augmented dickey fuller's (ADF) test was also applied to determine if the series was stationarity. The table below shows the results of the ADF test for both the infection and mortality cases.

Table 2: Augmented dickey fuller's test

Series	adf test	Lag	p-value
infections	0.146	6	0.99
mortality	-2.079	6	0.543

The p-value for the ADF test was higher than the alpha value (.05) therefore, the null hypothesis which stated that the series was non-stationary was not rejected implying the series was non-stationary. The following methods were applied to achieve stationarity.

4.4.1 Detrend Process

This is the process of removing the trend in the series to make the series stationary. Figure 7 and 8 below show the plots of the detrend infection and mortality cases respectively. It was observed that stationarity was not achieved through the detrending process.

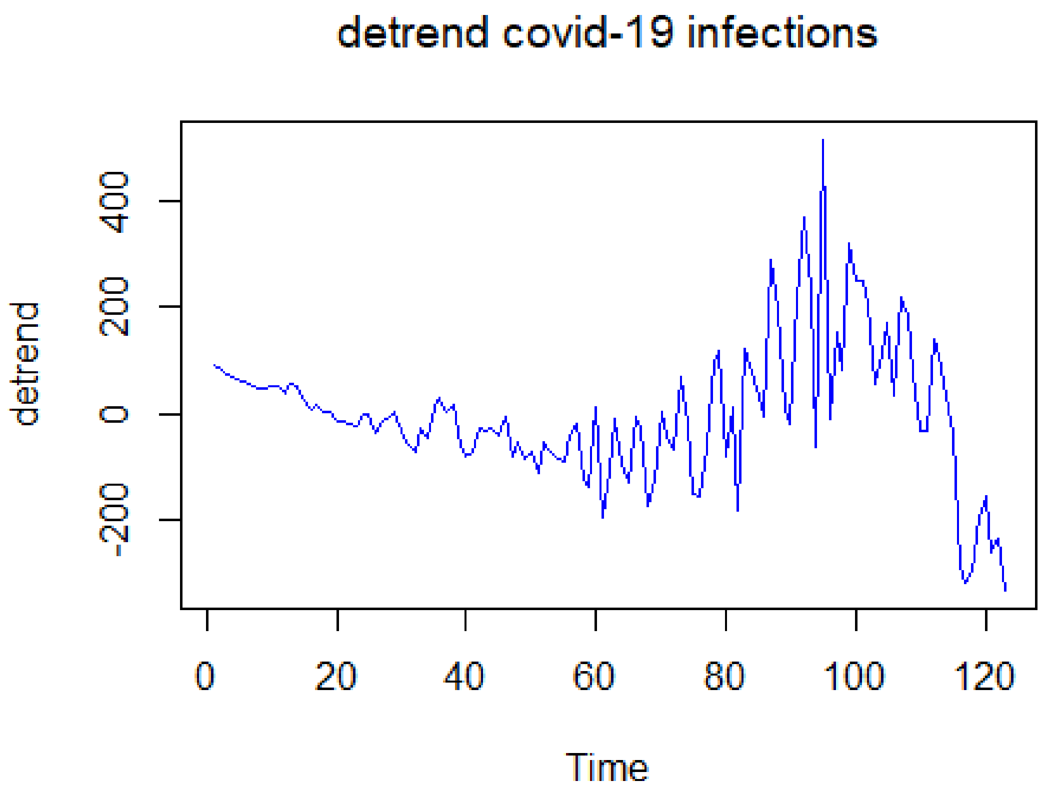

Figure 7: Plot of detrend COVID-19 infections

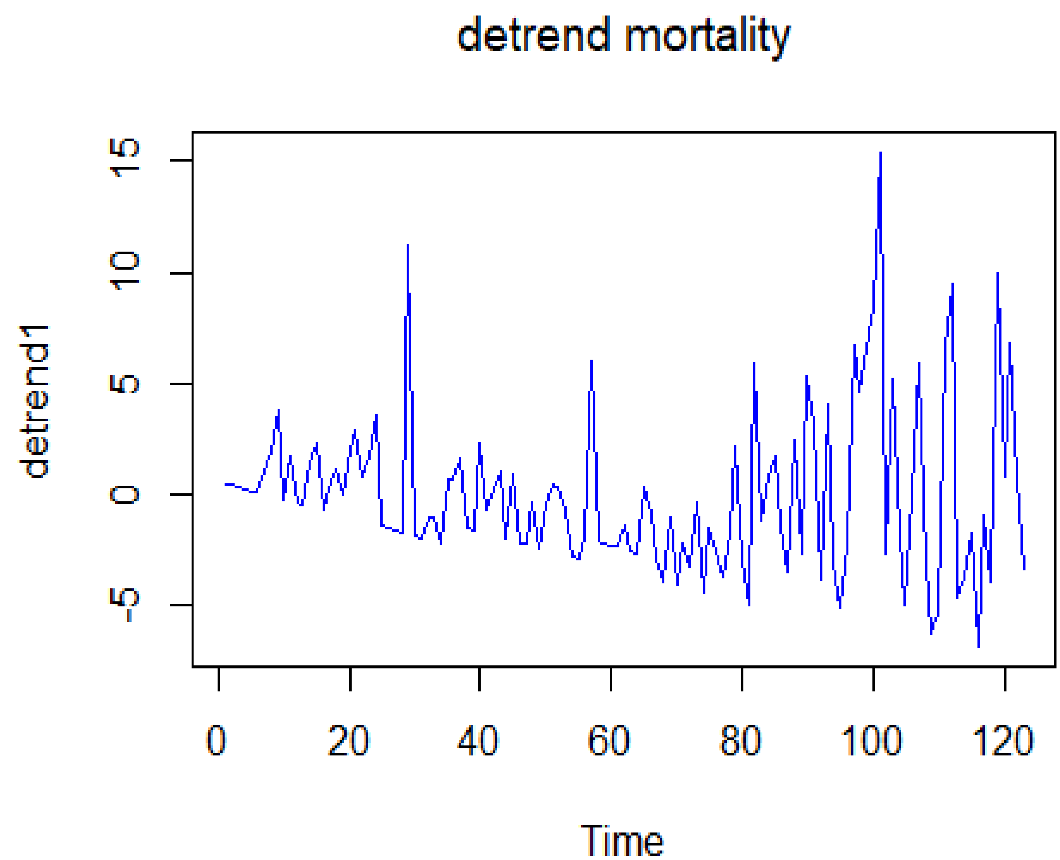

Figure 8: Plot of detrend COVID-19 mortality

4.4.2 Differencing Process

Differencing is the process of stabilizing the mean of the series by eliminating the trend and seasonality. The ndiffs function in R statistical software was applied to determine the order of differences that made the series stationary. After applying the ndiffs to the COVID-19 data, it was observed that stationarity was attained at first differencing. Figure 9 and 10 shows the plot of differenced COVID-19 infections and mortality respectively.

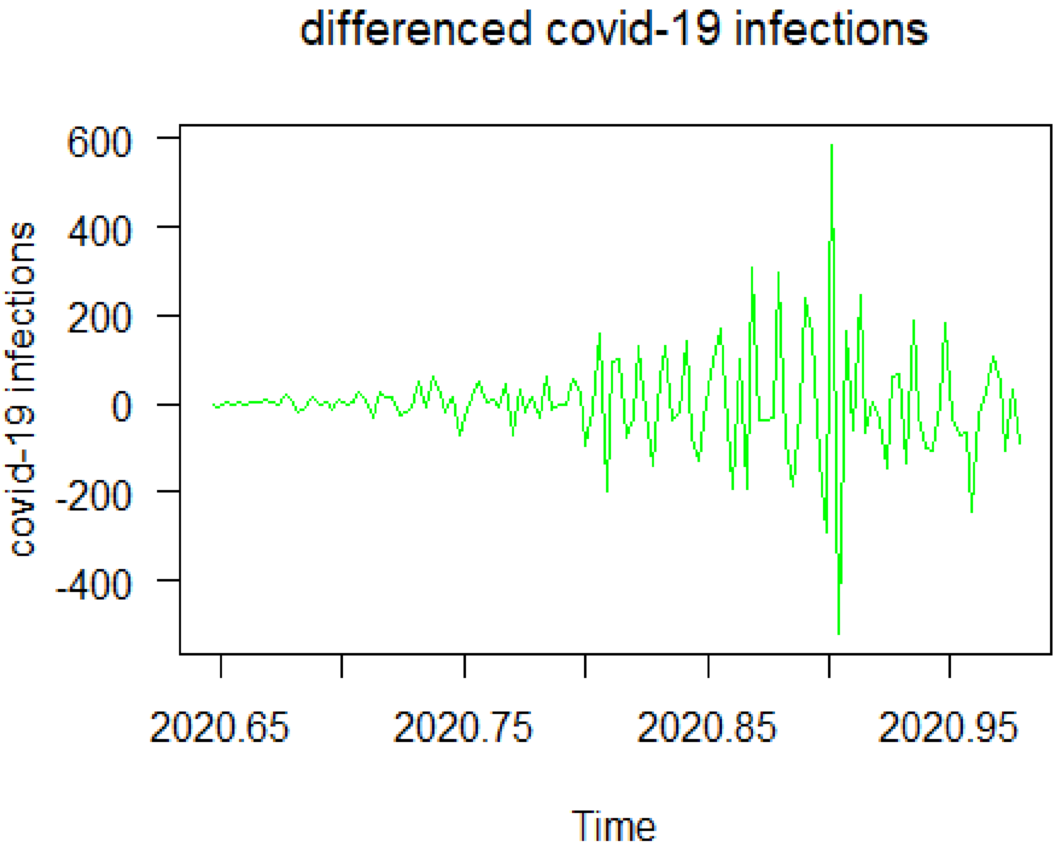

Figure 9: Plot of differenced COVID-19 infections

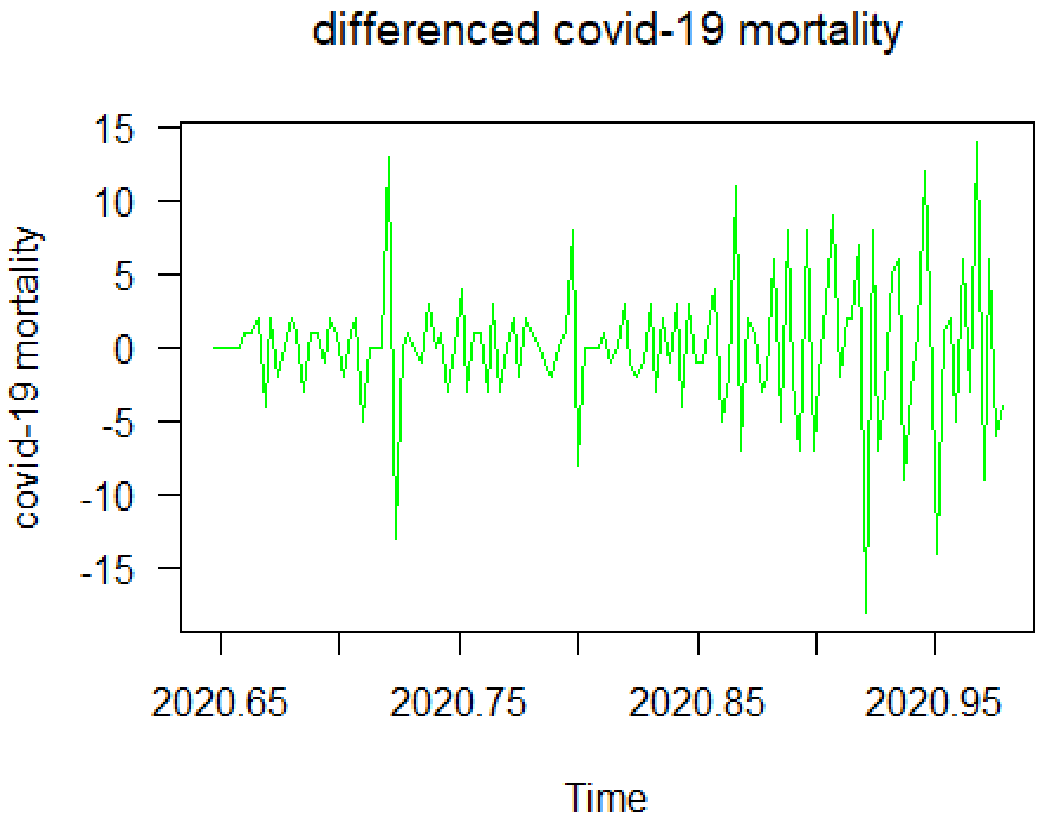

Figure 10: Plot of differenced COVID-19 mortality

4.4.3 ACF and PACF plots

The ACF plot and The PACF plot are used to determine the AR and the MA order of a given series. From the ACF and PACF plot, the spikes lie outside the significant zone (dotted line) implying that the residuals are not random and therefore, more information is needed to determine the AR order and the MA order. Figures 11, 12, 13 and 14 below displays the ACF and the PACF plots for the differenced COVID-19 infections and mortality respectively.

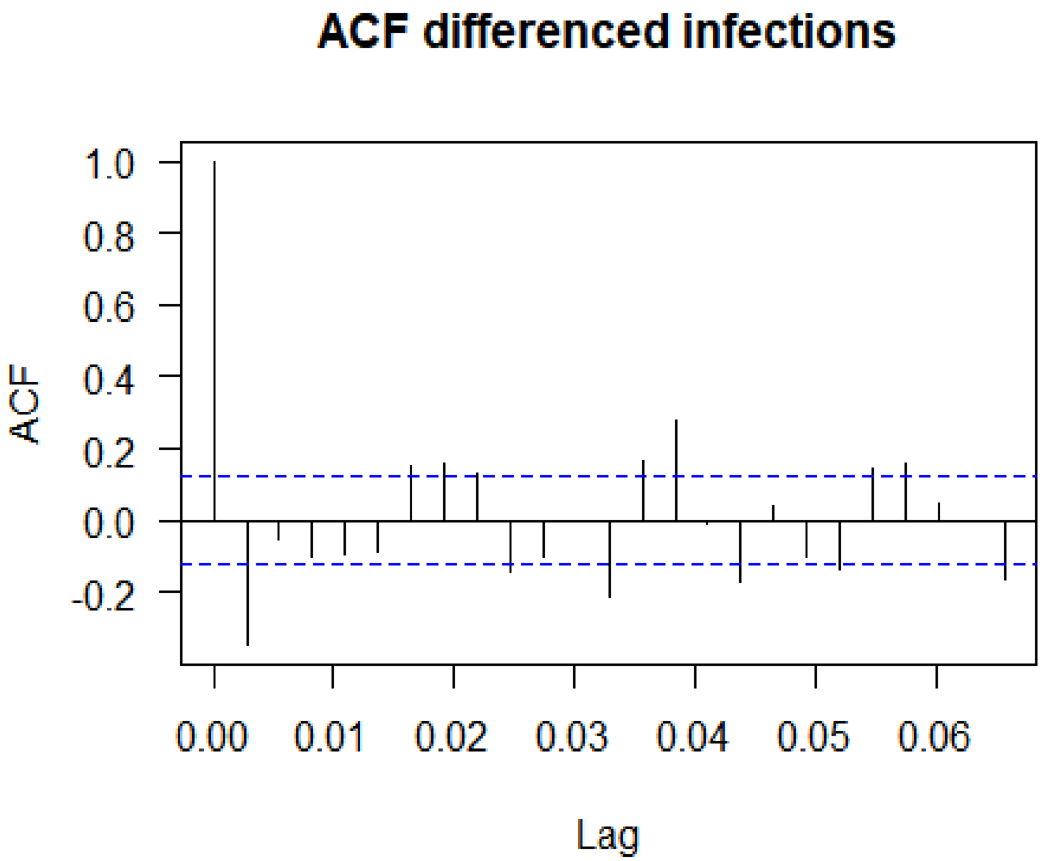

Figure 11: ACF plot of the differenced COVID-19 infections

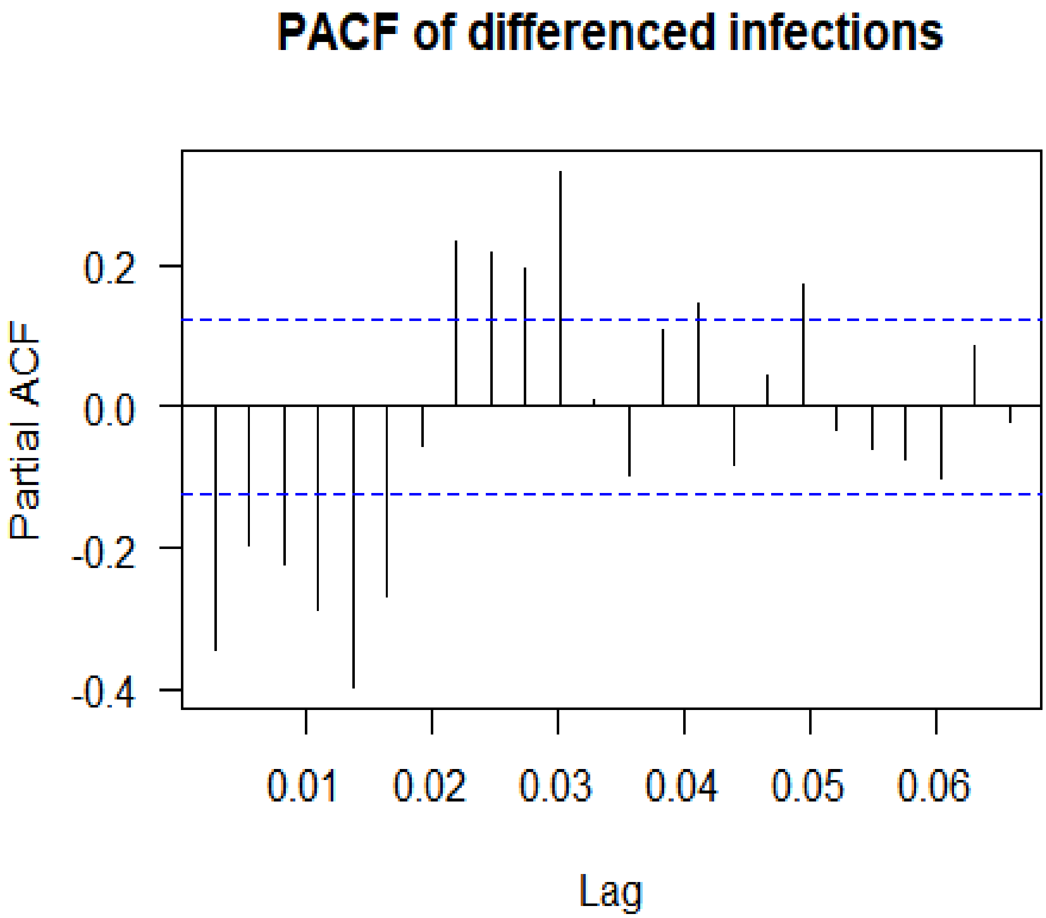

Figure 12: PACF plot of the differenced COVID-19 infections

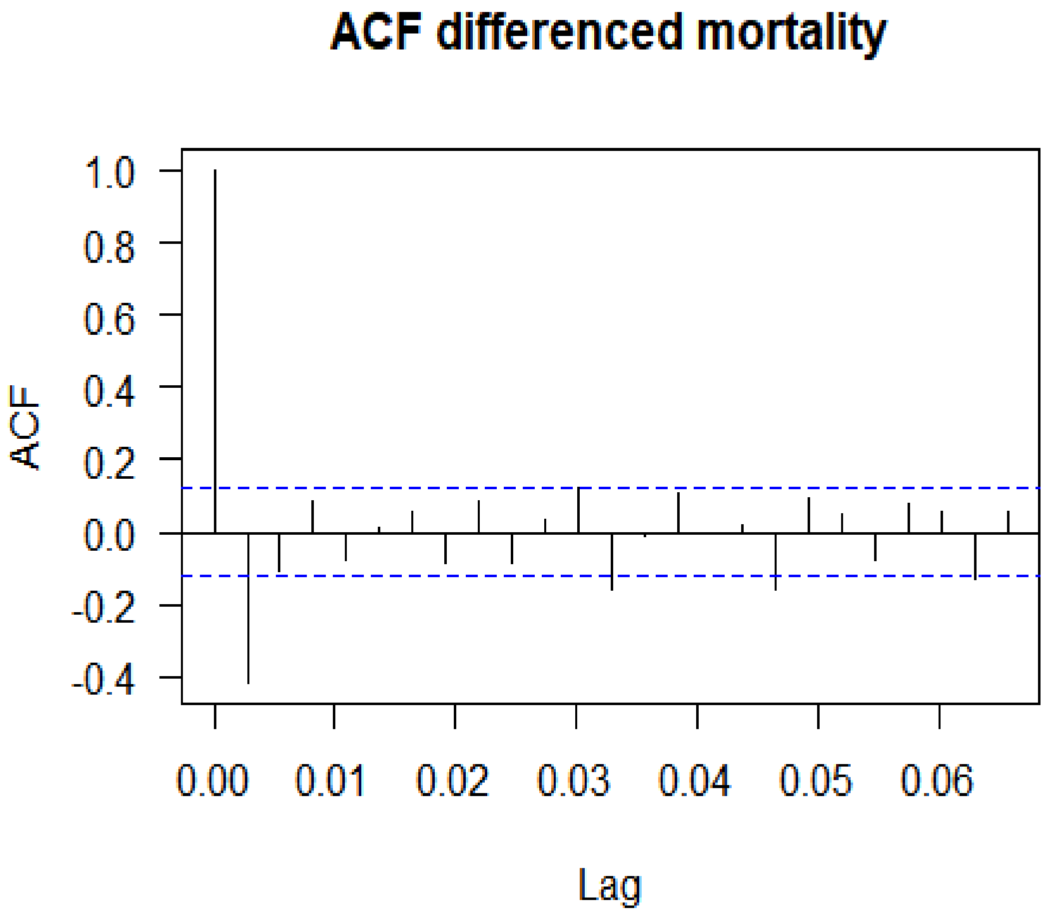

Figure 13: ACF plot of the differenced COVID-19 mortality

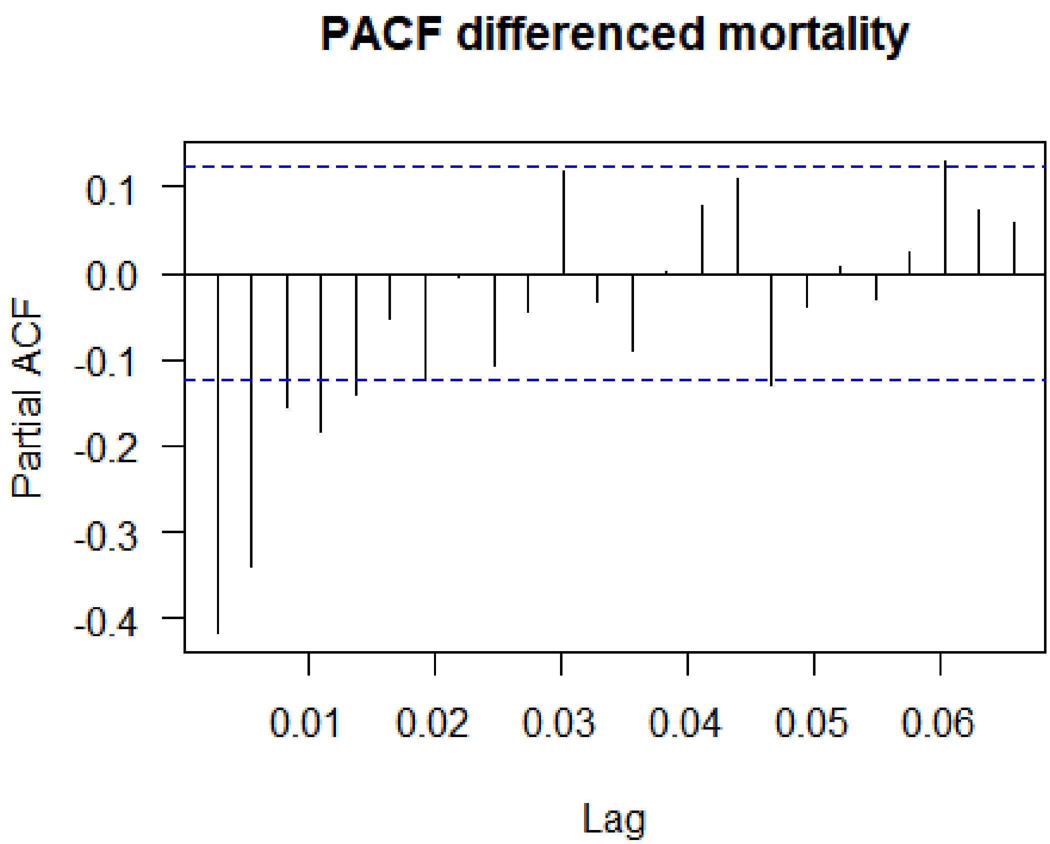

Figure 14: PACF plot of the differenced COVID-19 mortality

4.5 ARIMA Modeling

4.5.1 Model Selection and Parameter Estimation

To identify the AR and MA order, the ARIMA model was fitted in R software using the auto.arima() which is function in the forecast package in R software. The ARIMA(2,1,3) and ARIMA(0,1,1) for the COVID-19 infections and mortality were identified as the best models with the Akaike Information Criterion (AIC) values of 338.99 and 1406.05 respectively. Based on the two models, the AR value is 0 implying the absence of the autoregressive process in the series. The MLE estimation technique was applied to estimate the parameters in the model and the estimated parameters were, $\hat{\sigma}^2 = 0.2208$ and the log-likelihood = -166.5 for the ARIMA(2,1,3) and $\hat{\sigma}^2 = 15.28$ and the log-likelihood = -701.03 for the ARIMA(0,1,1) respectively. Table 3 shows the ARIMA(2,1,3) and the ARIMA(0,1,1) and their coefficients. From the tables, it was observed the coefficients for the ARIMA(2,1,3) are significant at a 5% except for the AR(1) whereas the coefficient for the ARIMA(0,1,1) is significant at a 5% since its p-value was less than alpha (.05).

Table 3: T-table for the ARIMA(2,1,3)

Coefficients	Estimate	Std. Error	t value	p-value
ar1	0.1618	.1821	0.8884	0.3761
ar2	-0.9169	0.1112	-8.2477	0.0000
ma1	-0.8330	0.2339	-3.5605	0.0005
ma2	0.9720	0.1869	5.2001	0.0000
ma3	-0.5635	0.1215	-4.6382	0.0000
Constant	2.6760	2.9397	0.9103	0.3646

Table 4: T-table for the ARIMA(0,1,1)

Coefficients	Estimate	Std. error	t value	p-value
ma1	-0.924	0.0454	-20.3377	0.0000
Constant	0.0794	0.0283	2.8091	0.0058

The regression equation for ARIMA(2,1,3) was expressed as

$$z_t = 2.6760 - 0.9169e_t - 0.8330e_{t-1} + 0.9720e_{t-2} - 0.5635e_{t-3}$$

whereas, the regression equation for the ARIMA(0,1,1) was expressed as

$$z_t = 0.0794 - 0.924\varepsilon_t$$

4.6 Holt-Winters Exponential Smoothing

The Holt-Winters models were fitted in R software using the HoltWinters function which is an inbuilt function in the forecast package in the R software. The differenced COVID-19 infection and mortality data was applied for the Holt-Winters modelling. The smoothing parameters and the coefficients were estimated as; α = 0.08, β=0.7, a = 237.77, b = -28.27 for the differenced COVID-19 infection and α = 0.09, β=0.02, a = 9.93 and b = 0.09 for the differenced COVID-19 mortality respectively. The absence of γ implies that there is no seasonality component in the series.

4.7 Model Comparison

The Root Mean Square Error (RMSE) was applied for model comparison. After devolving the models, the accuracy measures were determined. Table 5 and 6 shows the accuracy measures for the ARIMA model and Holt-Winters model for both the COVID-19 infection and mortality.

Table 5: Accuracy measure for the ARIMA model

Model	ME	RMSE	MAE
infections	0.529	0.0467	0.352
mortality	4.043	3.894	5.917

Table 6: Accuracy measure for Holt-Winters model

Model	ME	RMSE	MAE
infections	1.018	167.699	109.026
mortality	7.045	5.909	2.940

Based on the RMSE on the tables above, the ARIMA(2,1,3) and the ARIMA(0,1,1) were found to be more accurate prediction models with minimum errors compared to the Holt-Winters exponential smoothing method.

4.8 Diagnostic Test

The residuals were checked to determine whether they were independent and identically distributed. This was achieved by plotting the residuals plots. From the residual plots, the linearity of the Q-Q plot suggested that the model residuals were normally distributed. According to the ACF plots in Figures 17 and 19, it was observed that the majority of the spikes lie within the significant zone (dotted line) implying that the residuals are random.

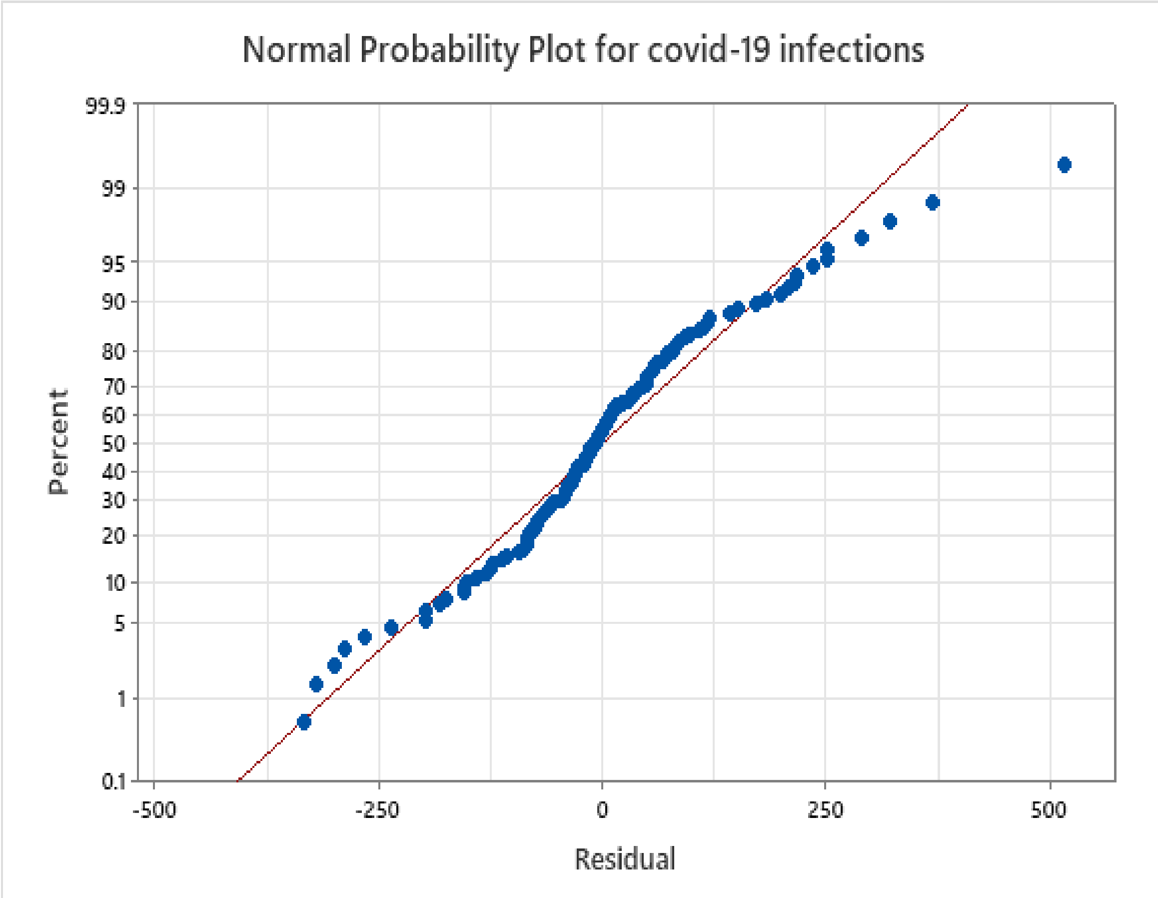

Figure 15: Normal Q-Q plot for the residuals

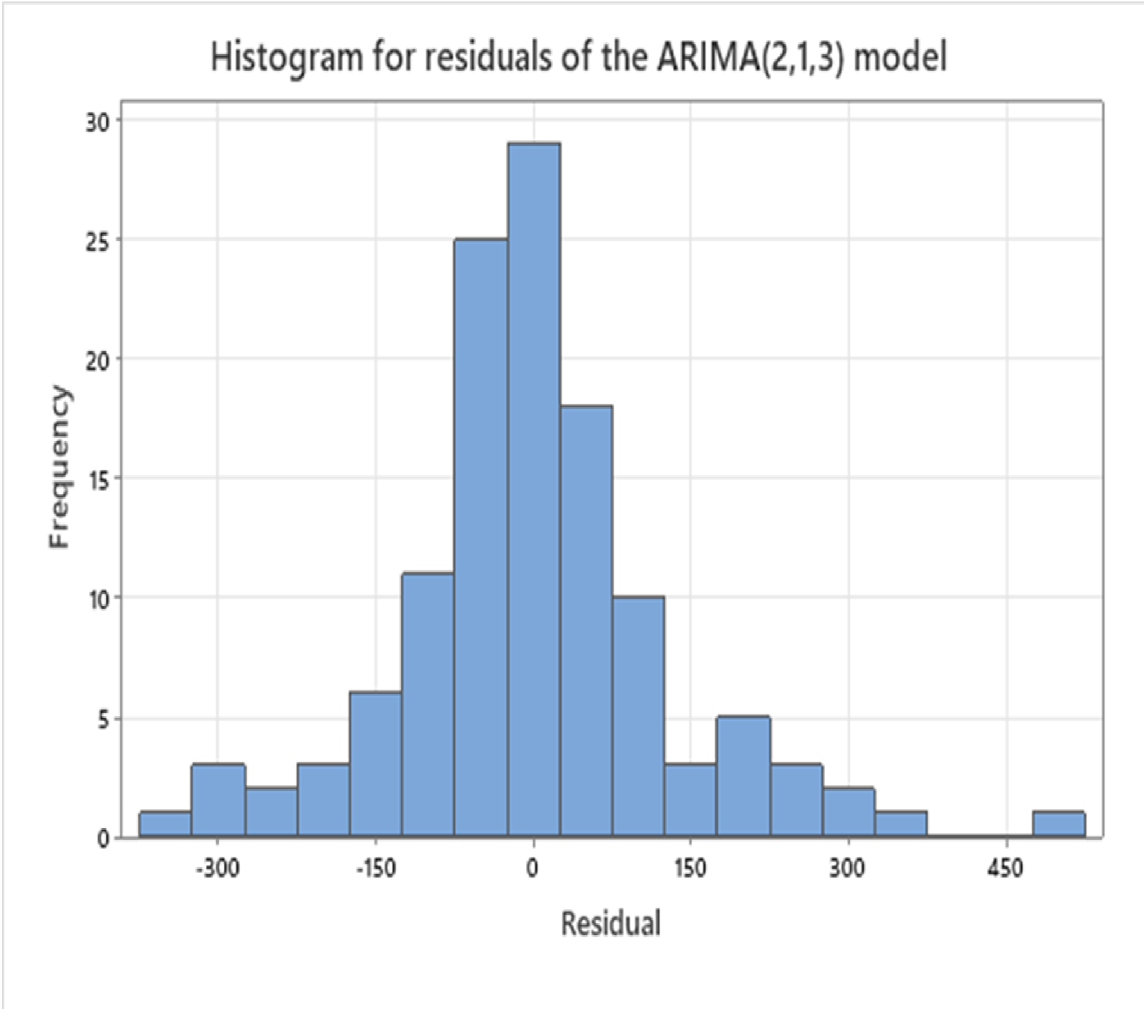

Figure 16: Histogram plot for the ARIMA (2,1,3)

The bell-shaped histogram in Figures 16, 18 and 19 implies that the model residuals of were normally distributed because the majority of the values are concentrated at the origin.

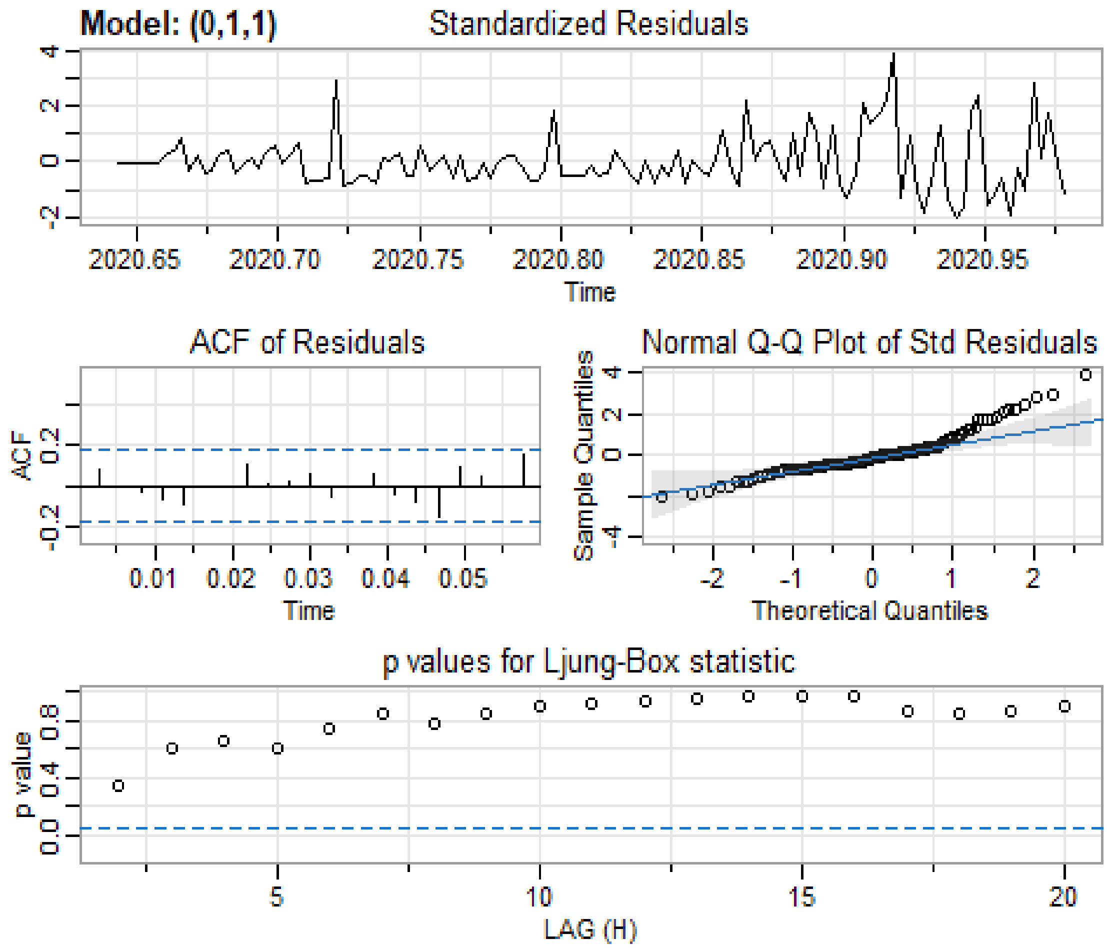

Figure 17: Residual plots for the ARIMA(0,1,1)

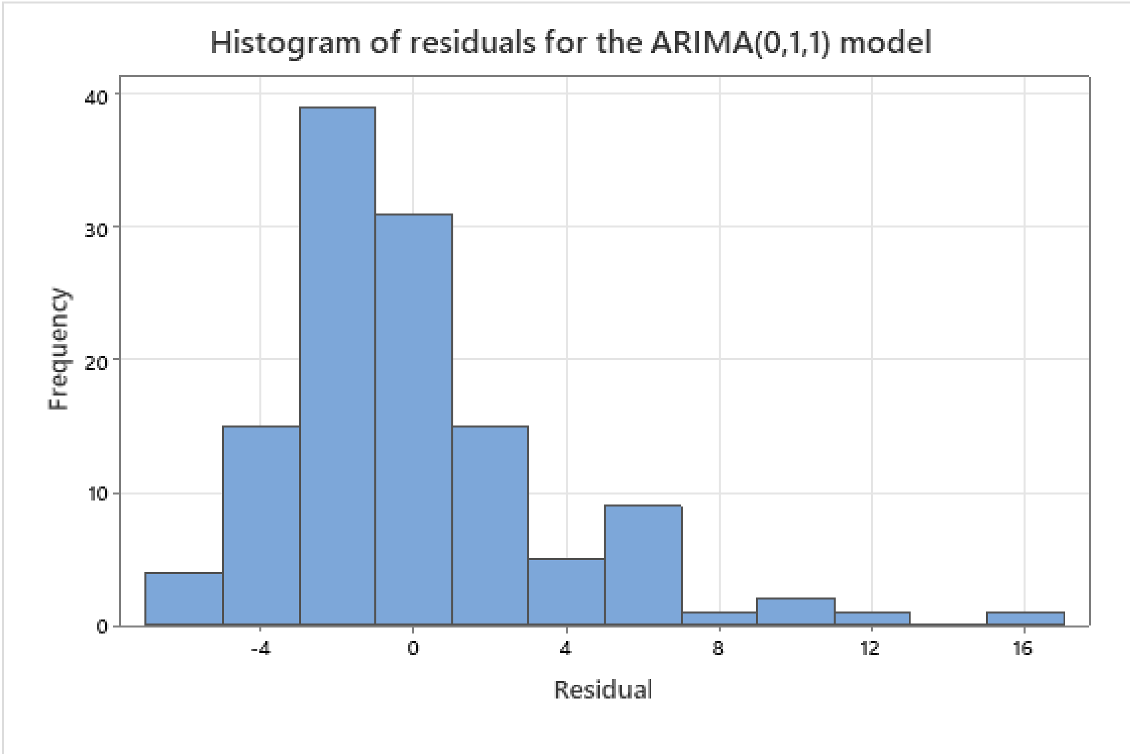

Figure 18: Histogram for the ARIMA(0,1,1)

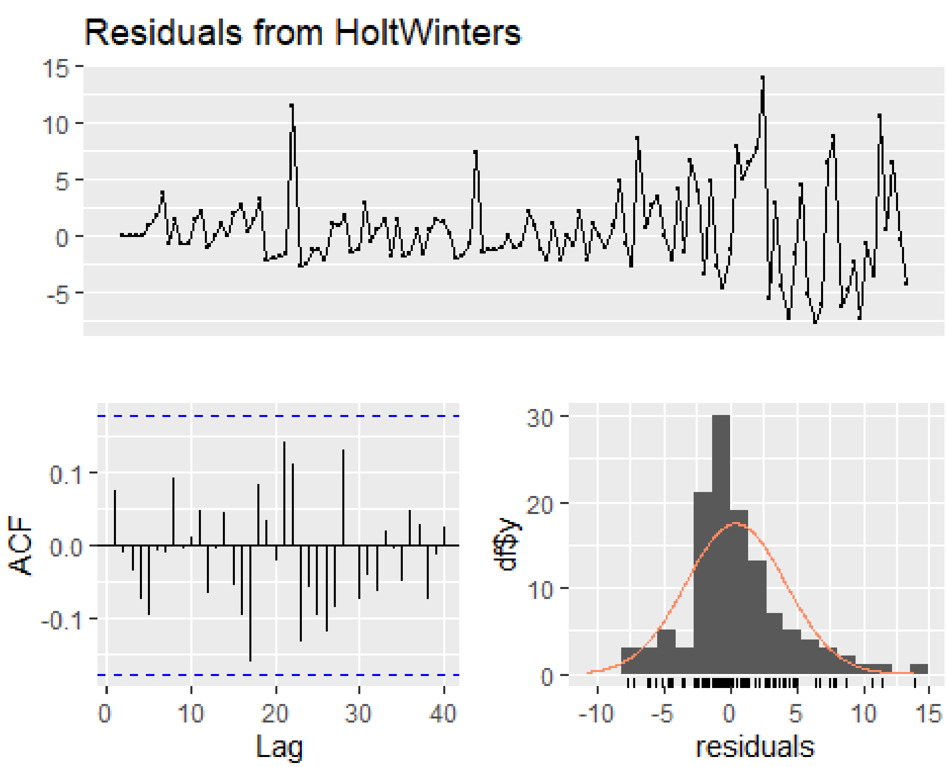

Figure 19: Residual plots for Holt-Winters models

The p-value for the Box Ljung test statistics is greater than the alpha value (.05) implying a non-significance therefore, we fail to reject the null hypothesis which stated that there was no serial correlation between the lags and conclude that the residuals were distributed as white noise. Since the ARIMA(2,1,3) and the ARIMA(0,1,1) models passed the diagnostic test, they were therefore applied to predict the COVID-19 infections and mortality cases in Kenya. The table below displays the results for the Box Ljung test for the ARIMA(2,1,3), the ARIMA(0,1,1) and Holt-Winters respectively.

Table 7: Ljung-Box test

Model	$\boldsymbol{Q}^{*}$	df	p-value
ARIMA for COVID-19 infections	16.429	20	0.6285
ARIMA for COVID-19 mortality	21.427	20	0.5489
Holt-Winters for COVID-19 infections	35.453	20	0.1782
Holt-Winters for COVID-19 infections	10.633	20	0.9552

4.9 Forecasting

The prediction for the COVID-19 infections and mortality were done at a 95% significance level using the ARIMA and the Holt-Winters models. The prediction was done for 92 days.

The figures below show the prediction plots for the COVID-19 infections and mortality respectively using the Holt-Winters and the ARIMA models. Based on the prediction plots below, the COVID-19 prediction using the ARIMA models was more accurate and realistic compared to the prediction using the Holt-Winters models. Therefore, the

ARIMA models were considered to provide a better prediction accuracy for the COVID-19 infections and mortality than the Holt-Winters models.

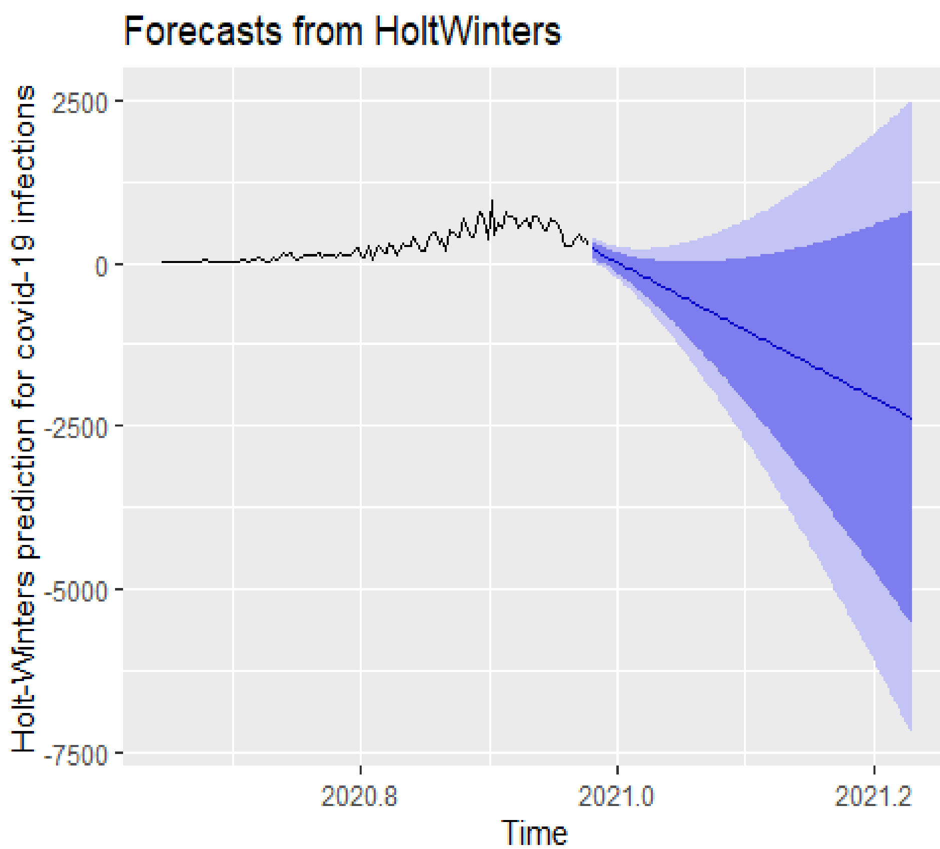

Figure 20: Holt-Winters prediction for COVID-19 infections

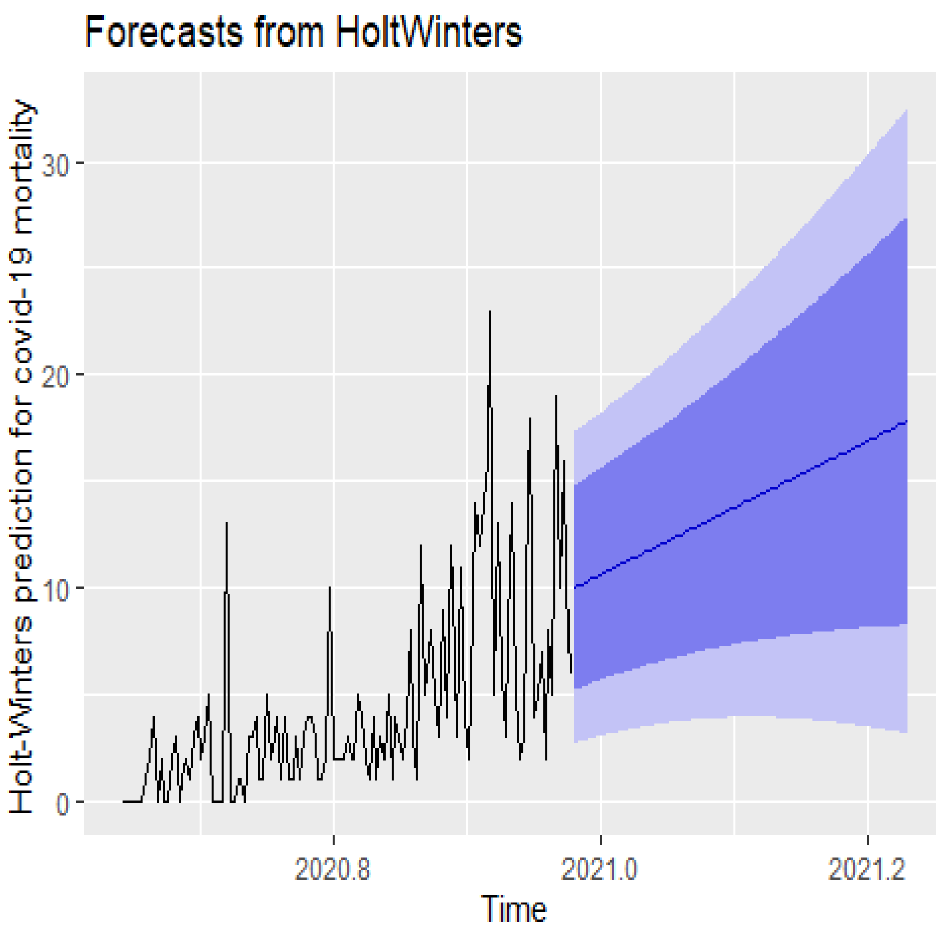

Figure 21: Holt-Winters prediction for COVID-19 mortality

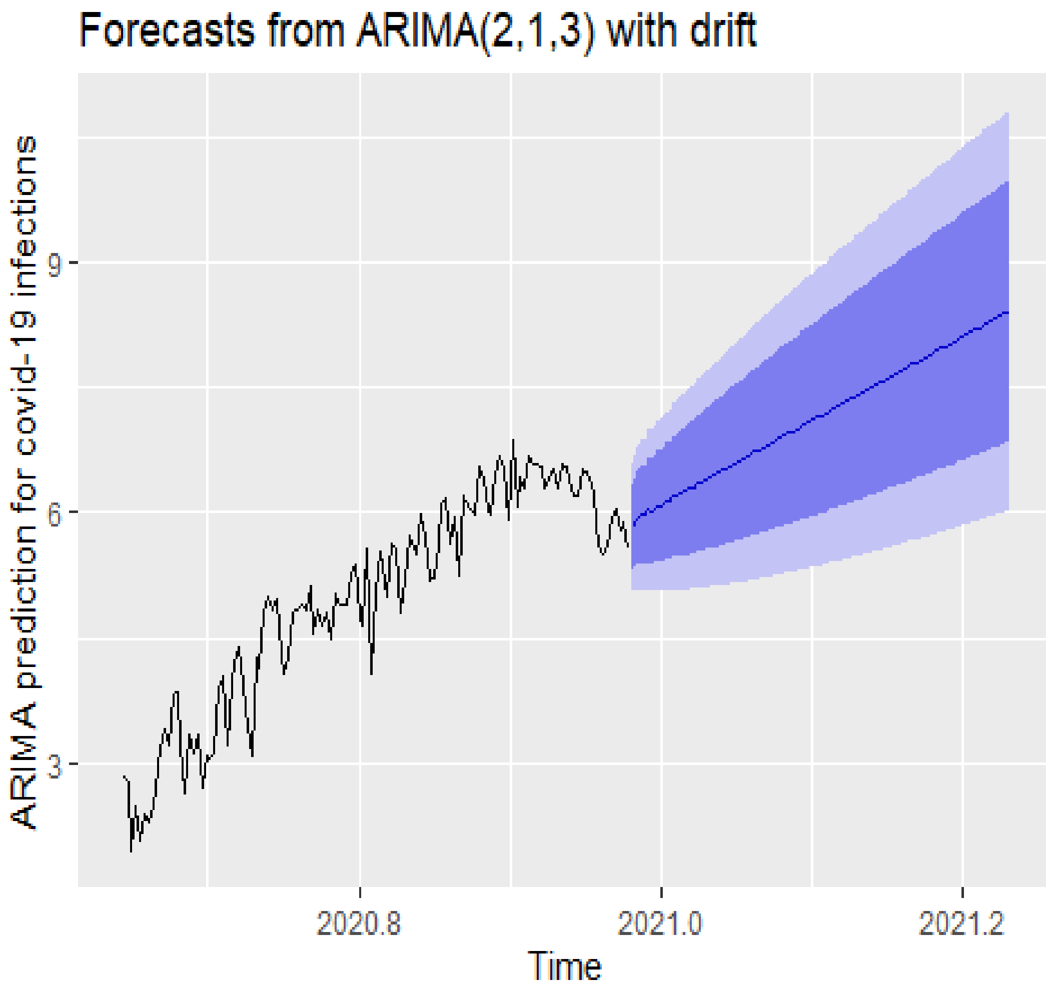

Figure 22: ARIMA prediction for COVID-19 infections

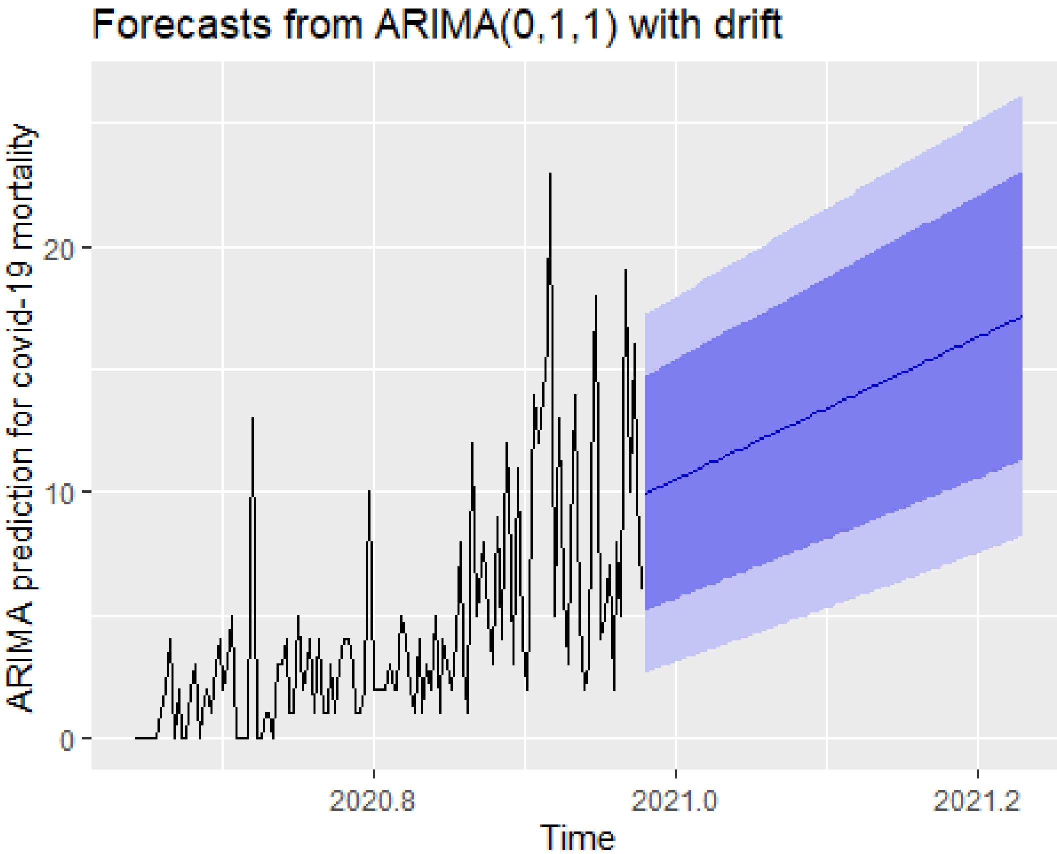

Figure 23: ARIMA prediction for COVID-19 mortality

CHAPTER FIVE

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.1 Discussion

This study developed the ARIMA and the Holt-Winters models to predict the COVID-19 infections and mortality in Kenya. It therefore compared the prediction accuracy of the ARIMA and the Holt-Winters models to determine an effective model for predicting the COVID-19 infection and mortality rates in Kenya. A simple linear regression was conducted to examine the linear relationship between the number of infected cases and the daily samples tested. The regression coefficient between the number of infected cases and the daily samples tested was -99.170 and 0.109 respectively. The response variable was statistically significant at a 5% confidence level implying a linear relationship. therefore, the null hypothesis which stated that there was no linear relationship between the number of infected cases and the daily samples tested was rejected implying the existence of a linear relationship between the two variables. The R-Squared was 0.72 implying that 72% of the variation in the response variable is explained by the explanatory variable. The scatterplot was also plotted to visualize the linear relationship. A linear relationship between COVID-19 infections and the sample size was observed.

Trend Analysis was conducted on the COVID-19 infection and mortality rates to determine the presence of trend in the series and a linear trend was identified. Since the COVID-19 infections and mortality time series data were not stationary, differencing was done to achieve the stationarity. The models were fitted to the data using R statistical software and the models with the lowest Akaike Information Criterion were selected. The Root Mean Square Error (RMSE) was applied for model comparison between the ARIMA and the Holt-Winters models to determine an accurate prediction model. The ARIMA(2,1,3) and the ARIMA(0,1,1) were considered better prediction models for COVID-19 infections and mortality than the Holt-Winters models since, the RMSE for the ARIMA models were much lower compared to that of Holt-Winters models implying that the ARIMA models yielded accurate prediction results with minimum errors. The diagnostic test was done on the fitted models and the models passed the test. The prediction for the COVID-19 infections and mortality were done at a 95% significance level and the infections and mortality cases were observed to increase significantly. Therefore, the ARIMA models predicted the COVID-19 infections and mortality perfectly well. The findings of this study contradicts with the findings of Veiga et al. (2014) who did a study on the prediction of food demand in retail by comparing the prediction accuracy of the ARIMA and the Holt-Winters models. Their study suggested that the Holt-Winters model was more precise than the ARIMA model. On the other hand, findings of this study are similar to that of Barr et al. (2021), Mini et al. (2015) and Lidiema (2017) who suggested that the ARIMA model was more accurate than the Holt-Winters model.

5.2 Conclusion

Based on the regression analysis, the study concluded that there was significant linear relationship between the sample size and the number of positive cases recorded daily. The study developed the ARIMA and the Holt-Winters models and the models to predict the COVID-19 infections and mortality perfectly well. The ARIMA and the Holt-Winters models were compared using the Root Mean Square Error (RMSE) and it was concluded that the ARIMA was a better prediction model for COVID-19 infections and mortality than the Holt-Winters due to minimum errors. Therefore, all the objectives of this study were achieved.

5.3 Recommendations

With the increasing COVID-19 infections and mortality rates in Kenya, the study recommends that all Kenyans should observe the World Health Organization's guidelines to help curb the surging COVID-19 infections and mortality rates. The government of Kenya should plan accordingly towards fighting the pandemic by provide affordable face masks to all Kenyans, and personal protective equipment for the medical personnel. Based on the prediction accuracy between the ARIMA and the Holt-Winters models, the study recommends that the ARIMA model be applied for short-term prediction because it is more accurate that the Holt-Winters model.

REFERENCES

Barbastefano, R., Carvalho, D., Clara, M., & Haime, D. (2020). A novel predictive mathematical model for COVID-19 pandemic with quarantine , contagion dynamics , and environmentally mediated transmission. *MedRxiv*, *6308*, 0–1.

Barr, C., Ferreira, G., Benz-parra, K., & Pablo, L. (2021). Prediction of confirmed and death cases of COVID-19 in Chile through time series techniques : A comparative study. *MedRxiv*.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID- 2019 epidemic dataset. *Data in Brief*, *29*, 105340.

Bhattacharya, A., & Banerjee, T. (2020). COVID-19 PANDEMIC IN INDIA : A STATISTICAL APPROACH. *International Research Journal of Modernization in Engineering Technology and Science*, *02*(07), 615–622.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2013). Time series analysis: Forecasting and control: Fourth edition. In *Time Series Analysis: Forecasting and Control: Fourth Edition*.

Chintalapudi, N., Battineni, G., & Amenta, F. (2020). COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy : A data driven model approach. *Journal of Microbiology, Immunology and Infection*, *53*(3), 396–403.

Da Veiga, C. P., Da Veiga, C. R. P., Catapan, A., Tortato, U., & Da Silva, W. V. (2014). Demand forecasting in food retail: A comparison between the Holt-Winters and ARIMA models. *WSEAS Transactions on Business and Economics*, *11*(1), 608–614.

Davydenko, A., & Fildes, R. (2013). Measuring Forecasting Accuracy: The Case Of Judgmental Adjustments To Sku-Level Demand Forecasts. *International Journal of Forecasting*, *29*(3), 510–522.

Gathungu et. al. (2020). Modeling the Effects of Non-Pharmaceutical Interventions on COVID-19 Spread in Kenya. *Interdisciplinary Perspectives of Infectious Diseases*, *2020*, 1–15.

Gelper, S., Fried, R., & Croux, C. (2010). Robust Forecasting with Exponential and Holt–Winters Smoothing. *Journal of Forecasting*, *300*(June 2009), 285–300.

Habibur Rahman, M., Salma, U., Moyazzem Hossain, M., & Tareq Ferdous Khan, M. (2016). Revenue Forecasting using Holt–Winters Exponential Smoothing. *Research & Reviews: Journal of Statistic*, *5*(3), 19–25. www.stmjournals.com

Keith, W., & Mcleod, A. (1994). Capter 7 Diagnostic Checking. *Development in Water Science*, *45*, 235–253.

Khrapov, P. V, & Loginova, A. A. (2020). Mathematical modelling of the dynamics of the Coronavirus COVID-19 epidemic development in China. *International Journal of Open Information Technologies*, *8*(4), 2307–8162.

Kumar, P., Singh, R. K., Nanda, C., Kalita, H., Patairiya, S., Sharma, Y. D., Rani, M., & Bhagavathula, A. S. (2020). Forecasting COVID-19 impact in India using pandemic waves Nonlinear Growth Models. *MedRxiv*, 1–18.

Lidiema, C. (2017). Modelling and Forecasting Inflation Rate in Kenya Using SARIMA and Holt-Winters Triple Exponential Smoothing. *American Journal of Theoretical and Applied Statistics*, *6*(3), 161.

Mahmoud, E. (1984). Accuracy in forecasting: A survey. *Journal of Forecasting*, *3*(2), 139–159.

Makatjane, K. D., & Moroke, N. D. (2016). Comparative study of Holt-Winters Triple Exponential Smoothing and Seasonal ARIMA: Forecasting Short Term Seasonal car sales in South Africa. *Journal of Chemical Information and Modeling*, *6*(1), 1689–1699.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, *9*(4), 527–529. https://doi.org/10.1016/0169-2070(93)90079-3

Maleki, M., Mahmoudi, M. R., Wraith, D., & Pho, K. H. (2020). Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel Medicine and Infectious Disease*, *May*, 101742.

Maleki, M., Reza, M., & Hossein, M. (2020). Modeling and forecasting the spread and death rate of coronavirus (COVID-19) in the world using time series models. *Chaos, Solitons and Fractals*, *140*, 110151.

Mathenge Mutwiri, R. (2019). Forecasting of Tomatoes Wholesale Prices of Nairobi in Kenya: Time Series Analysis Using Sarima Model. *International Journal of Statistical Distributions and Applications*, *5*(3), 46.

Maurice, W. (2019). *Modelling Factors Affecting Lung Capacity*. *34*(6), 1–18.

Mini, K. G., Kuriakose, S., & Sathianandan, T. V. (2015). *Modeling CPUE series for the fishery along northeast coast of India : A comparison between the Holt- Winters , ARIMA and NNAR models*. *57*(12).

Mohammad, F., & Gupta, R. (2020). Journal of Safety Science and Resilience 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.

Odhiambo, J. O., Ngare, P., Weke, P., & Otieno, R. O. (2020). *Modelling of COVID-19 Transmission in Kenya Using Compound Poisson Regression Model*. *35*(2), 101–111.

Omane-adjepong, M., Oduro, F. T., & Oduro, S. D. (2013). Determining the Better Approach for Short-Term Forecasting of Ghana's Inflation: Seasonal-ARIMA vs Holt-Winters. *International Journal of Business, Humanitiies and Technology*, *3*(1), 69–79.

Pandey, G., Chaudhary, P., Gupta, R., & Pal, S. (2019). *SEIR and Regression Model based*

COVID-19 outbreak predictions in India. 1–10.

Rajagopal, K., Hasanzadeh, N., Parastesh, F., Ismael, I., Jafari, S., & Hussain, I. (2020). A fractional-order model for the novel coronavirus (COVID-19) outbreak. *Nonlinear Dynamics*.

Singh, S., Singh, K., Kumar, J., Jitendra, S., & Makkhan, S. (2020). Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19. *Chaos, Solitons and Fractals: Nonlinear Science, and Nonequilibrium and Complex*, *135*, 1–8.

Takele, R. (2020). Stochastic modelling for predicting COVID-19 prevalence in East Africa Countries. *Infectious Disease Modelling*, *5*, 598–607.

Tinani, K., Muralidharan, K., Deshmukh, A., & Patil, B. (2020). Analysis and Forecasting of COVID-19 Cases Across Hotspot States of India. *Statistics and Applications*, *18*(1), 223–238.

Wang, H., Wang, Z., Dong, Y., Chang, R., Xu, C., Yu, X., & Zhang, S. (2020). Phase-adjusted estimation of the number of Coronavirus Disease 2019 cases in Wuhan , China. *Cell Discovery*, *6*(1), 1–8.

Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons and Fractals: The Interdisciplinary Journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena*, *139*, 110058.

WHO. (2020). WHO siterep 73. *World Health Organization*, *2019*(3), 2633. https://doi.org/10.1001/jama.2020.2633

Xiao, Y., & Wu, J. (2020). An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov). *Infectious Disease Modelling*, *5*, 248–255.

Yakovyna, V., & Bachka, O. (2018). The comparison of holt – winters and box – jenkins methods for software failures prediction. *CEUR Workshop Proceedings*, *2136*, 90–98.

Yang, Q., Wang, J., Ma, H., & Wang, X. (2020). as an example to see the epidemic in Italy. *Journal of Infection and Public Health*, 4–7.

Zhao, B., Wang, Z., Yu, Z., Tian, C., & Cao, J. (2020). Time Series Analysis of the Novel Coronavirus (COVID-19). *Immunology and Allergy*, *1*(3), 1–13.

Zhao, Z., Li, X., Liu, F., Zhu, G., Ma, C., & Wang, L. (2020). Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria. *Science of the Total Environment*, *729*, 138959.

Zhou, X., Ma, X., Hong, N., Su, L., Ma, Y., & He, J. (2020). *Forecasting the Worldwide Spread of COVID-19 based on Logistic Model and SEIR Model*.

APPENDICES

Appendix 1: R codes for data analysis

```
library(readxl)

setwd("C:/Users/HP/Desktop/R data Files")

COVID.19=read_excel('ResearchData.xlsx')

attach(COVID.19)

head(COVID.19,5)

data=ts(COVID.19[,2],frequency = 365,start=c(2020,236))

data30=ts(COVID.19[,3],frequency = 365,start=c(2020,236))

plot(data,main="plot of COVID-19 infections",ylab="COVID-19
infections",las=1,col="red",font.main=1)

plot(data30,main='plot of COVID-19 mortality',col='red',las=1,ylab="mortality
cases",font.main=1)

library(lmtest)

trend=lm(Registered~c(1:length(Registered)))

detrend=residuals(trend)

plot.ts(detrend,main='detrend COVID-19 infections',col='blue',font.main=1)

trend1=lm(Fatalities~c(1:length(Fatalities)))

detrend1=residuals(trend1)

plot.ts(detrend1,main='detrend mortality',col='blue',font.main=1)

library(tseries)

adf.test(data,alt="stationary")

adf.test(data30,alt="stationary")
```

```
plot(diff(data),main="differenced COVID-19 infections",las=1,ylab="COVID-19 infections",col="green",font.main=1)

plot(diff(data30),main="differenced COVID-19 mortality",las=1,ylab="COVID-19 mortality",col="green",font.main=1)

acf(diff(data),main="Correlogram",las=1,font.main=1)

pacf(diff(data),main="Partial correlogram",las=1,font.main=1)

acf(diff(data30),main="Correlogram",las=1,font.main=1)

pacf(diff(data30),main="Partial correlogram",las=1,font.main=1)
```

#ARIMA Model Estimation for COVID-19 infections

```
library(forecast)

ARIMAfit=auto.arima(log(data))

summary(ARIMAfit)

library(astsa)

sarima(data,2,1,3,font.main=1)

library(lmtest)

coeftest(ARIMAfit)
```

#ARIMA model Estimation for COVID-19 mortality

```
Modelfit=auto.arima(data30)

summary(Modelfit)

library(astsa)

sarima(data30,0,1,1,font.main=1)

library(lmtest)

coeftest(Modelfit)
```

#Holt-Winters model estimation for COVID-19 infections

```
HWfit=HoltWinters(data,gamma=F)

print(HWfit)

library(forecast)

pred1=forecast(HWfit,h=92)

accuracy(pred1)
```

#Holt-Winters model estimation for COVID-19 Mortality

```
HWmodel=HoltWinters(data30,gamma=F)

print(HWmodel)

library(forecast)

pred2=forecast(HWmodel,h=92)

accuracy(pred2)
```

#Diagnostic test for the COVID-19 infections

```
checkresiduals(ARIMAfit,main="ARIMA residuals",font.main=1)

checkresiduals(HWfit,main="Holt-Winters residuals",font.main=1)
```

#Diagnostic test for the COVID-19 mortality

```
checkresiduals(Modelfit,main="ARIMA residuals",font.main=1)

checkresiduals(HWmodel,main="Holt-Winters residuals",font.main=1)

Box.test(Modelfit$residuals,lag=20)

Box.test(ARIMAfit$residuals,lag=20)

Box.test(pred1$residuals,lag=20)

Box.test(pred2$residuals,lag=20)
```

#ARIMA Prediction

```
fcast1=forecast(ARIMAfit,h=92)

fcast2=forecast(Modelfit,h=92)

library(ggplot2)

autoplot(fcast1,ylab='ARIMA prediction for COVID-19 infections',font.main=1)

autoplot(fcast2,ylab='ARIMA prediction for COVID-19 mortality',font.main=1)

autoplot(pred1,ylab='Holt-Winters prediction for COVID-19 infections',font.main=1)

autoplot(pred2,ylab='Holt-Winters prediction for COVID-19 mortality',font.main=1)
```