ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

Modelling Daily COVID-19 Cases in Kenya Using ARIMA Model

Caroline M. Kamotho[1],Josephine N. Ngure[1], Margaret W. Kinyua[2]

[1]*Department of Pure and Applied Sciences, Kirinyaga University, P.O. Box 143-10300, Kerugoya (Kenya),*

[2]*Department of Mathematics Statistics and Actuarial Science,Karatina University, P.O. Box 1957-10101, Karatina (Kenya)*

DOI: https://doi.org/10.51584/IJRIAS.2023.8621

Received: 01 June 2023; Revised: 16 June 2023; Accepted: 22 June 2023; Published: 15 July 2023

Abstract: Severe Acute Respiratory Syndrome is the primary cause of the current pandemic coronavirus disease (COVID-19). The first case was reported in Wuhan, China, on December 30th, 2019 with the first case on 13thMarch, 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. Researchers have fitted time series models but using a short data length and without a transition. There was therefore a need to model a longer data period of daily COVID-19 cases with a transition in Kenya using theAutoregressiveIntegrated Moving Average (ARIMA) model and forecast. Secondary data from the World Health Organization from 13thMarch, 2020 to 30thApril, 2023 was analyzed using R software. The data was found to be non-stationary using the Augmented Dickey Fuller test and regular differencing was done to make it stationary. The Box-Jenkins methodology was used to fit the model of the data and afterwards forecasting was done. The ARIMA (3,1,2) was selected as the best model since it had the least Akaike Information Criterion and Bayesian Information Criterion among the possible models. Model validation using test data was done by comparing the MAE, and RMSE of the model's forecasts and it was the best amongst the possible models with MAE = 2.77 and RMSE =2.88. The model was fitted to the daily COVID-19 data and forecasting was then done for ninety days into the future.

KeyWords:COVID-19,ARIMA,trend,seasonality, forecasts

I. Introduction

The novel coronavirus disease, named COVID-19 by the WHO on February 11th, 2020, is a contagious disease caused by a Severe Acute Respiratory Syndrome Coronavirus 2 known as the SARS-CoV-2 virus. This virus is from the large family of coronavirus (CoVs) which cause Severe Acute Respiratory Syndrome (SARS).

The name SARS-CoV-2 was adopted after the genetically related SARS-CoV by the International Committee on Virus Taxonomy on February 11th, 2020. WHO uses COVID-19 to refer to SARS-CoV-2 to avoid confusion with the SARS disease which sounds almost the same.PolymeraseChain Reaction (PCR) reversetranscriptase is used to confirm the presence of the virus in an individual. Most of the infected people experience minimal to moderate symptoms and fully recover with no treatment. However, some experience severe symptoms and require medical consultation. Older people and those having underlying conditions have a higher chance of developing this illness but any individual can be infected with COVID-19 despite their age. Symptoms appear in five to six days on average, but it can take up to fourteen days for a person to become infected with the virus [12].

Frequent mild signs and symptoms include a high fever, cough, exhaustion, and loss of taste or smell while other mild symptoms include; sore throat, headache, aches and pains, diarrhea, a rash on the skin, discoloration of fingers or toes, and red or irritated eyes [22]. Patients with severe symptoms like breathing difficulty or shortness of breath, loss of speech or mobility and chest pain are advised to consult a doctor. The constant evolution 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) [10].

The virus can spread from an infected person's mouth in minute droplets [11]. Then one can catch the virus by breathing them in or by touching their eyes, nose, or mouth after touching a contaminated surface. This virus spreads in enclosed spaces and more rapidly in open spaces especially when the virus is in the air and there is wind blowing [7].

The number of confirmed positive cases were usually recorded on a daily basis and made available for researchers to analyze and find its characteristics as secondary data [23]. Time series models are usually fitted to data recorded over a period of time. Time series data is said to be stationary when there is no systematics change in mean and variance or non- stationary when it contains trend, seasonality, cyclic effect or a combination of any of these components.

When using non-stationary data, differencing is done to make the data stationary which results to an ARIMA model. The general ARIMA model is simply the ARMA model including the differencing part to make the time series data stationary.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

Researchers have fitted most of time series models to the COVID-19 data then forecast the COVID-19 cases but they can be easily affected by overfitting. ARIMA model is the most fitted model but the aims of different researchers might be different.

Globally, the COVID-19 pandemic is one of the most dangerous diseases to world public health, posing an unsettling scenario with more than six million deaths. The negative impacts of COVID-19 did not spare the Big 4 agenda and the SDG's in Kenya. In previous studies, most researchers analysed the COVID-19 data then forecasted the disease cases using ARIMA models in Kenya. However, data over a long period of time with a transition point has not been analysed using an ARIMA model in Kenya.

Forecasting is simply using historical COVID-19 data to predict or estimate the future values. In this study, the ARIMA model is fitted to the daily COVID-19 data with an aim of estimating the best model that will fit the data then use the model to forecast for future 90 days.

II. Method And Analysis

A. Source of Data

The target population comprised of the total number of COVID-19 positive cases in Kenya which was recorded from March 2020 to April 2023. These secondary data was sourced from the WHO's website and was analyzed using R software.

B. ARIMA Model Formulation

*1) 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 Xt is said to be an auto-regressive process of order p denoted by AR(P) if

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} \quad (1)$$

2) Moving Average (MA) Model: The MA model, also known as the MA process is a basic time series model which is finitely stationary and is mostly used to model univariate time series data. We can generally say that it is a linear combination of present and past values of a white noise error term. A process Xt is said to be a moving average process of order q denoted by MA(q) if

$$X_t = \beta_0 e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + ... + \beta_q e_{t-q} \quad (2)$$

3) Autoregressive Moving Average (ARMA) Model: This is a mixture of two models, the AR and MA models. 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) where p and q are the orders of the AR and MA models respectively [6]. It can be written as

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + ... + \beta_q e_{t-q} \quad (3)$$

4) Stationarity: A time series is said to be stationary if it has no systematic change in the mean i.e no trend, no systematic change in variance (homoscedasticity) and if there are no periodic variations. One way to check for stationarity is by observing the time plot and the correlogram [9]. The ADF is also used to test the stationarity of data and it test ensures to reject the null hypothesis (the time series is not stationary) since it assumes that the process is not stationary. For this to be achieved, the p-value should be less than the level of significance hence the inference to be made will be that the process is stationary.

5) Differencing: If a time series is not stationary, it can be made stationary by differencing. This was done by subtracting one value from another successive value once to achieve stationarity. According to [14], 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.

6) Autoregressive Integrated Moving Average (ARIMA) Model: To remove the parameters causing non-stationarity we perform differencing [24]. An ARMA model when subjected to differencing becomes an ARIMA (p, d, q) model where (p, d, q), explains the order of the model. According to[18], ARIMA (p, d, q) model is used on the non-seasonal data to predict future values based on past observations only.

Where;

p is the order of the AR model

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

d is the order or the number of differencing

q is the order of MA model.

The model is

$$W_t = \alpha_1 W_{t-1} + \alpha_2 W_{t-2} + ... + \alpha_p W_{t-p} + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + ... + \beta_q e_{t-q} \quad \textbf{(4)}$$

Where $W_t = \nabla(d)X_t$is the difference

C. Box Jenkins Methodology

This method was majorly used for model identification, estimation and prediction [8]. The procedure was as follows;

1) Model Identification:

i.Data Preparation: A time series plot was plotted and data transformation using logarithm was performed to ensure variance stability. Test for stationarity was done and differencing was done once and data was stationary.

ii.Model Selection:During this stage, ACF and PACF were plotted to identify the appropriate models. The Akaike Information criterion (AIC) and the Bayesian Information criteria (BIC) estimates the quality of each model and the model with the least value was selected and assumed to be the best model.

Akaike Information Criterion: Given a set of models, it provides a means for selecting models as it estimates the quality of a model compared to the others. It ensures that the selected model is generalizable, and offers a balance between goodness of fit and parsimony.The model with the least AIC value is selected as the best model.

Bayesian Information Criterion*:* It was introduced by [19] 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.BIC is majorly used in finite models and the model with the lowest or the smallest BIC is considered to be the best one [16].

2) Parameter Estimation: Estimation of parameter values will be done using the Maximum Likelihood Estimator method [17]. This is because it is the most suitable and feasible method as there will be the error terms which can be classified as missing/unobserved data. Hence the likelihood can be easily obtained. Considering the daily COVID-19 cases, e.g X_1, X_2, X_3, ..., X_n is from a density function f (x, θ) where θ is the unknown parameter. The likelihood function was given by;

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta) \qquad (5)$$

*3) Diagnostic checking***:**In this step, there was checking and testing whether the model is adequate. IfthemodelwiththelowestAICandBICandalsohasnormallydistributedresiduals then it was selected as the best model. The residuals' autocorrelation was testedusing the Ljung-Box test in which if the p-value was less than 0.05, then thenull hypothesis that there is no autocorrelation was to be rejected.The Ljung-Box test is used to test the presence of autocorrelationintheresidualsofamodelhencetestingthegoodnessofamodel[20].

4) Forecasting: The best selected model was then to be used to forecast the future COVID-19 cases. In case the model was found to be inadequate in step 3 for some reasons, the researcher was then to proceed to construct and test the ARIMA model again until a satisfactory model is found.

D. Checking the Selected Model's Accuracy

Forecast accuracy also known as the forecast error is the difference between actual cases and forecasted cases [21]. The methods of Root Mean Squared Error and Mean Absolute Error were used to check the accuracy ofCOVID-19 forecasted number of cases because it is measured in the same units as the variables [5].They are calculated using the following formulae,

$$MAE = \frac{\sum_{i=1}^{p} \left| X_i - \hat{X}_i \right|}{p} \qquad (6)$$

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

$$RMSE=\sqrt{\frac{\sum_{i=1}^{p}(X_i-\hat{X}_i)^2}{p}} \quad (7)$$

Where,

X_i = the actual number of cases

$\hat{X}_i$= the forecasted number of cases

p= the number of observed COVID-19 cases

III.Results And Discussion

In this section, we discuss the results of the estimated ARIMA model and its accuracy in forecasting the model. The following is the time series plot of the data.

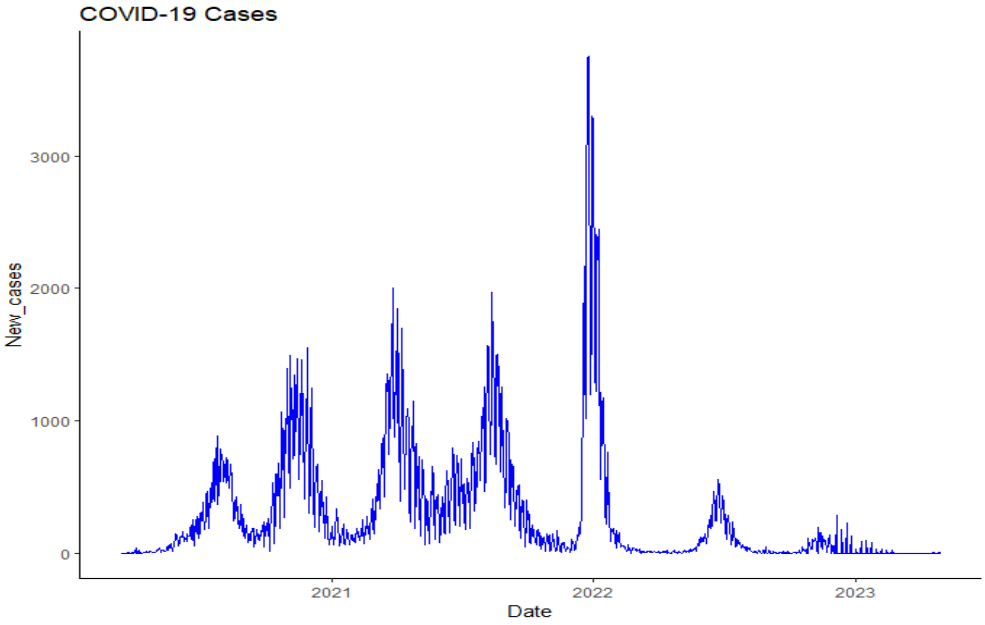

Fig. 1.Plot of COVID-19 cases.

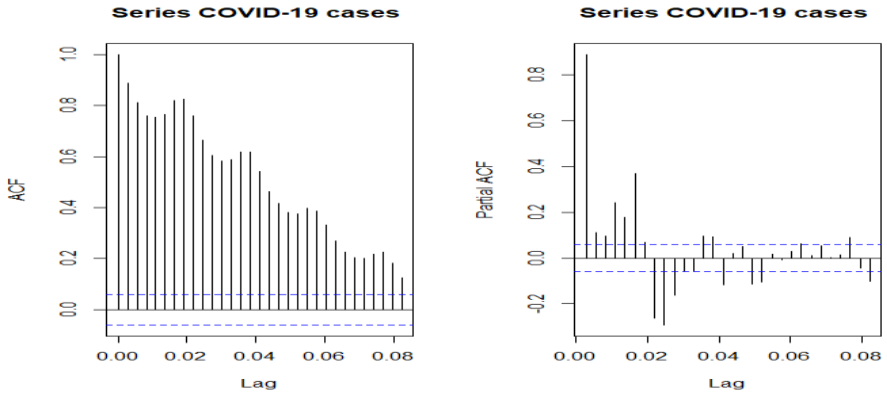

Fig. 2.ACF and PACF plot of COVID-19 cases.

The slow decaying pattern of the ACF showed that the time series was non-stationary. 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. Since the data revealed that it had different variances in Fig.1, log transformation was applied to the data to stabilize the variance.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

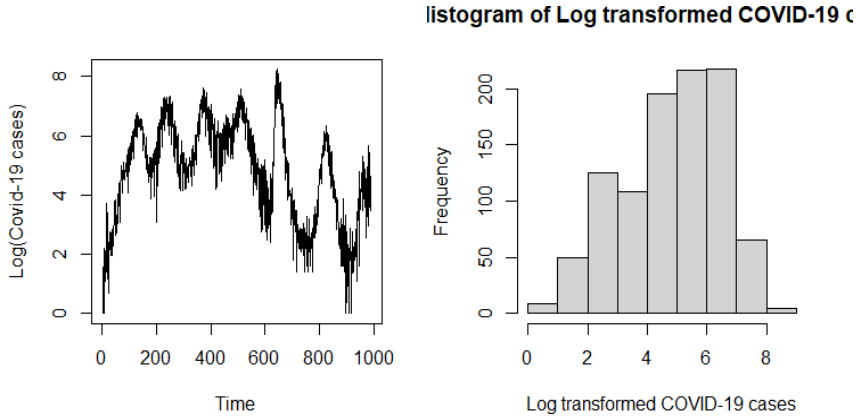

Fig. 3.Plot and Histogram of log-transformed COVID-19 cases.

Both the plot and the histogram showed that the variance of the data had been stabilized.

A. Testing for Stationarity

The ADF test was done to confirm whether the data was stationary hence the need for differencing or not. In the results, p-value = 0.1141 which was greater than 0.05 hence the null hypothesis that the data is stationary was not rejected indicating that the data was non-stationary and therefore differencing was required. Differencing was done and the following are the plots of the resulting log-differenced data.

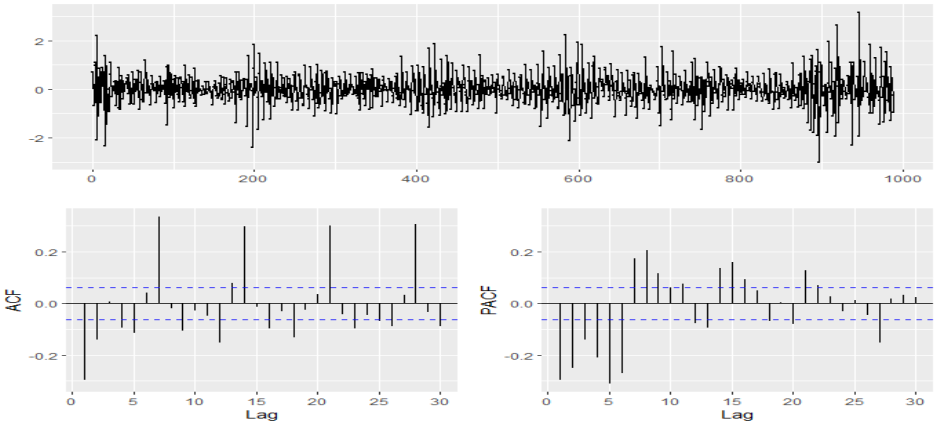

Fig. 4.Plot, ACF and PACF of log-differenced COVID-19 cases.

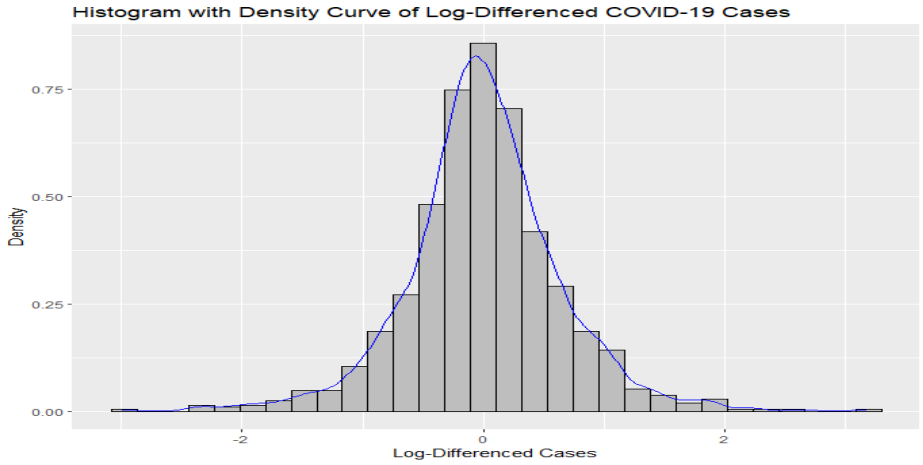

Fig. 5.Histogram of log-differenced COVID-19 cases.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

After differencing, the COVID-19 cases had a stable variance and constant mean. ADF test was carried out again to evaluate the data's stationarity and the result was a p-value = 0.01 hence the null hypothesis that the data was not stationary was rejected indicating that the data was stationary after 1st differencing.

B. Fitting the ARIMA Model

The model was fitted by generating the possible ARIMA models of the data picked from the ACF and PACF plots of the log-transformed data and the results were as shown in TableI.

Table I. Possible ARIMA models

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

ARIMA(3,1,2) was selected as the best model since it was the model with the least AIC, BIC and RMSE values. These results implied that the parameters of the ARIMA model were as follows,

p = 3 (order of AR model)

d = 1 (order of differencing)

q = 2 (order of MA model)

C. Diagnostic Checking

The residuals of the ARIMA (3,1,2) were tested using Ljung-Box test and the results were that the residuals had no auto-correlation hence independent from each other. This was deduced from the resulting p-value of 0.2561 which was greater than 5% level of significance which meant that there was no autocorrelation in the residuals as the null hypothesis that the residuals had no autocorrelation was not rejected.

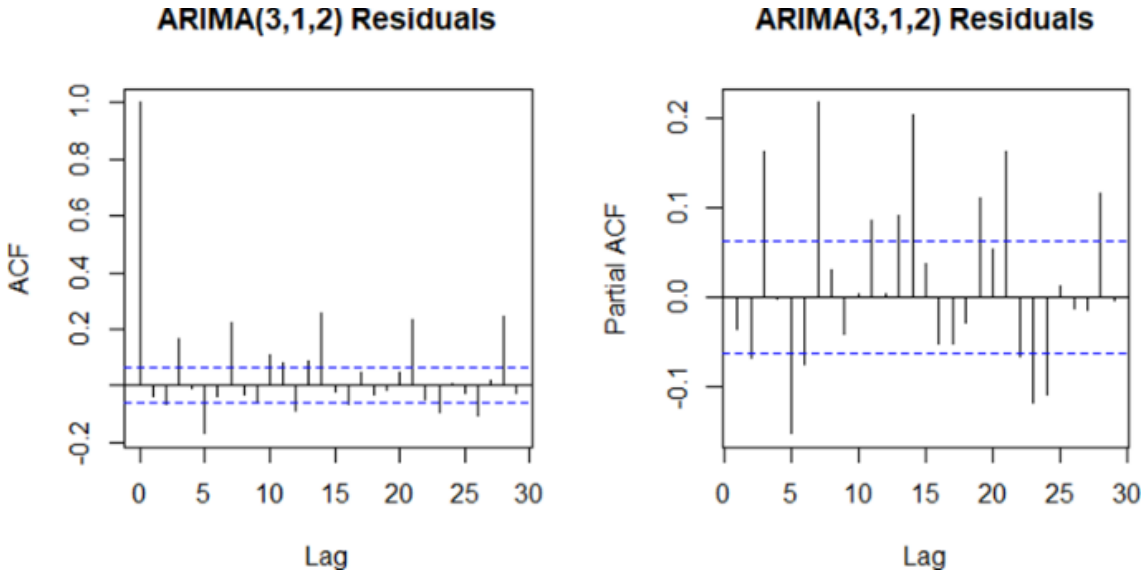

Fig. 6. ACF and PACF plots of ARIMA(3,1,2) residuals

The above figure is the ACF plot of the ARIMA(3,1,2) residuals. Although the ACF did not indicate perfect non autocorrelation, the Ljung-Box test proved that the there was no autocorrelation in the residuals.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

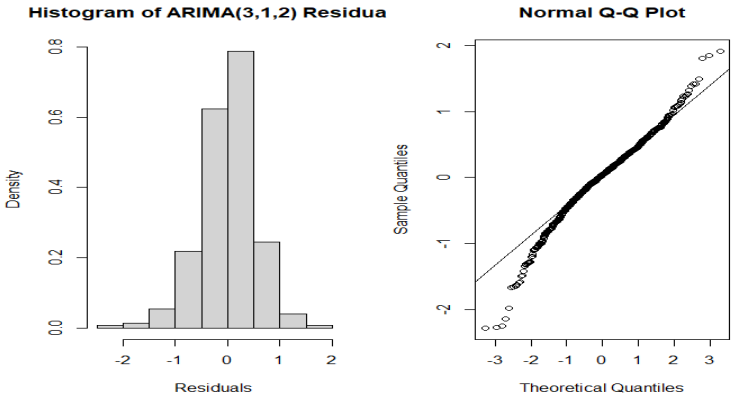

Fig. 7. Histogram and Q-Q plots of ARIMA (3,1,2) residuals

The histogram above and the Q-Q plot indicated that the residuals were normally distributed and hence confirmed that the model was a good fit to the data.

D. Model Validation

The back testing and eye testing tests were done to validate the model. Eye testing was done by observing the forecasts and they seemed reasonable compared to the actual values. Back testing was done using MSE and the RMSE of the forecast errors for the test data. The models were used to forecast for 51 dayswhich were then compared to 51 observations test dataset. The MAE, MSE and RMSE values of the different model's forecasts were compared as follows;

Table II. Model Validation Statistics

ARIMA Order	MAE	MSE	RMSE
(1,1,1)	3.0435	9.9932	3.1617
(1,1,2)	3.0477	10.0141	3.1645
(2,1,1)	3.0516	10.0339	3.1676
(2,1,2)	2.9179	9.1314	3.0218
(3,1,1)	3.0544	10.0458	3.1695
(3,1,2)	2.7690	8.3079	2.8823

From Table II above, the model ARIMA (3,1,2) was the model with the least validation statistics hence confirming that it was the best model that fit the data could be used for forecasting. The model was then written as;

$$X_t = 0.8317X_{t-1} - 0.3824X_{t-2} - 0.2377X_{t-3} + e_t - 1.3931e_{t-1} + 0.7084e_{t-2} \quad (7.1)$$

The model ARIMA (3,1,2) forecasts for 51 days against the test data set of 51 days are shown in the plot below.

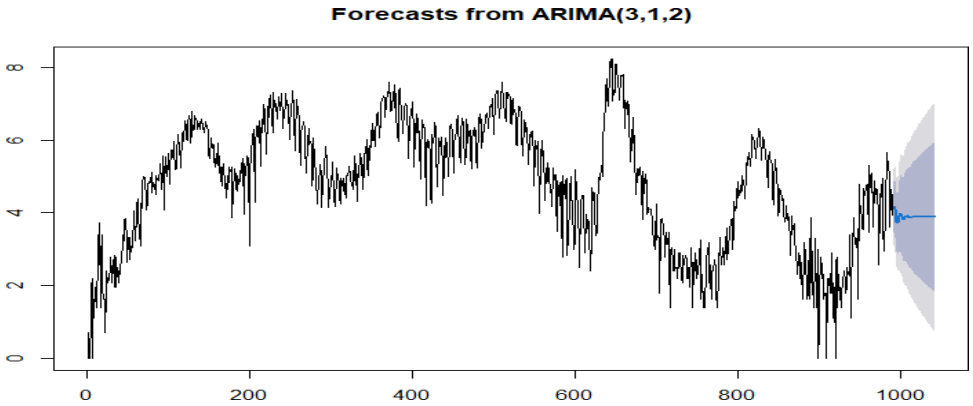

Fig. 8. Plot of ARIMA (3,1,2) forecasts for test dataset

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

E. Future Forecasting

The model was then used for future forecasting of 90 days and the plot of forecasts was as shown below.

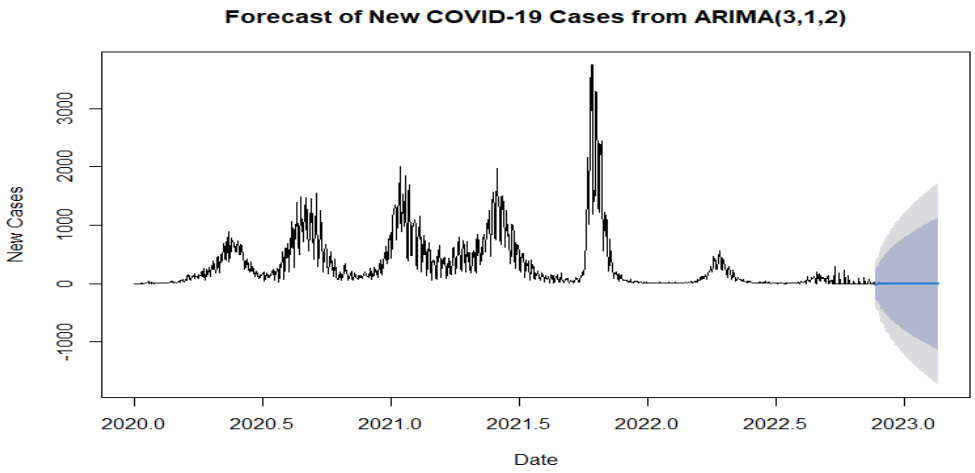

Fig. 9. Plot of ARIMA (3,1,2) future forecasts

IV. Conclusion

The COVID-19 data used had different variances data transformation was required to stabilize the variance, the differencing was also carried out to stabilize the mean. The estimated model was found as ARIMA (3,1,2) since it had the least AIC, BIC and RMSE values amongst the possible models. The model was then validated using MAE, MSE and RMSE of the forecast errors from the test set. The ARIMA (3,1,2) model was found to be the best model since its forecasts were closest to the actual COVID-19 cases. Therefore, it was used for future forecasting.

Due to the lifting of restrictions by the 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. The presence of two types of seasonality disqualified ARIMA time series modeling. One can use the BATS and TBATS models to model such data.

Acknowledgment

The author highly acknowledges the World Health Organization for providing the data used in this research. The author would also like to acknowledge Kirinyaga University for providing the requiredfacilityand faculty for guidance and supervision. TheDean of my department, Dr. Peter WanjohiNjori and also the Chair, Dr. Jeremiah Ndung'uKinyanjui for the support and patience as I worked towards finishing the project are highly appreciated.

References

1. Al Khames Aga, Q. A., Alkhaffaf, W. H., Hatem, T. H., Nassir, K. F., Batineh, Y., Dahham, A. T., Shaban, D., Al Khames Aga, L. A., Agha, M. Y., and Traqchi, M. (2021). Safety of covid-19 vaccines. Journal of medical virology, 93(12):6588– 6594.
2. Anderson, T. W. (2011). The statistical analysis of time series. John Wiley & Sons.
3. Brockwell, P. J. and Davis, R. A. (2009). Time series: theory and methods. Springer science & business media.
4. Cavanaugh, J. E. and Neath, A. A. (2019). The akaike information criterion: Back- ground, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics, 11(3):e1460
5. Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae). Geoscientific model development discussions, 7(1):1525–1534.
6. Choi, B. (2012). ARMA model identification. Springer Science & Business Media.
7. Cos¸kun, H., Yıldırım, N., and Gündüz, S. (2021). The spread of covid-19 virus through population density and wind in turkey cities. Science of the Total Environment, 751:141663.
8. Devi, B. U., Sundar, D., and 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.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VI June 2023

9. Dickey, D. A. (2015). Stationarity issues in time series models. SAS Users Group International, 30.
10. Duong, D. (2021). Alpha, beta, delta, gamma: What's important to know about sars- cov-2 variants of concern?
11. Jayaweera, M., Perera, H., Gunawardana, B., and Manatunge, J. (2020). Transmission of covid-19 virus by droplets and aerosols: A critical review on the unresolved dichotomy. Environmental research, 188:109819.
12. Lima, C. M. A. d. O. (2020). Information about the new coronavirus disease (covid- 19).
13. Márquez, F. P. G., Pedregal, D. J., and Roberts, C. (2015). New methods for the condition monitoring of level crossings. International Journal of Systems Science, 46(5):878–884.
14. Moh'dMussa, A. and Saxena, K. (2018). Trend analysis and forecasting of performance of students in mathematics in certificate secondary education examination in zanzibar: Arima modelling approach.
15. Mushtaq, R. (2011). Augmented dickey fuller test.
16. Neath, A. A. and Cavanaugh, J. E. (2012). The bayesian information criterion:background,derivation,andapplications.Wiley Interdisciplinary Reviews: Computational Statistics, 4(2):199–203.
17. 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, pages 1–24.
18. Prabhakaran, S. (2019). Arima model–complete guide to time series forecasting in python. Machine Learning Plus, 18.
19. Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, pages 461–464.
20. Serra, R. and Rodríguez, A. C. (2012). The ljung-box test as a performance indicator for vircs. In International Symposium on Electromagnetic Compatibility-EMC EUROPE, pages 1–6. IEEE.
21. 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.
22. Struyf, T., Deeks, J. J., Dinnes, J., Takwoingi, Y., Davenport, C., Leeflang, M. M., Spijker, R., Hooft, L., Emperador, D., Domen, J., et al. (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).
23. Vartanian, T. P. (2010). Secondary data analysis. Oxford University Press.
24. Zhang, M. (2018). Time series: Autoregressive models ar, ma, arma, arima. University of Pittsburgh.