Journal of Econometrics and Statistics
Vol. 3, Issue 2, 2023, pp. 185-200
https://doi.org/10.47509/JES.2023.v03i02.05

Modelling Daily New Cases of COVID-19 in Lagos State Nigeria. ARIMA or ARFIMA?

[1]Oluwagbenga Tobi Babatunde, [2]Chinaza Orji and [3]Abimibola Victoria Oladugba,

[1,2,3] Department of Statistics, University of Nigeria, Nsukka.

[1]Corresponding Author: oluwagbenga.babatunde@unn.edu.ng

Abstract: The impact of Coronavirus disease (COVID-19) is globally felt and understanding the spread or growth of the virus is one of the ways to flatten the curve of the virus. There is need to understand the spread of this virus in terms of future projections in other to put in adequate measures to curtail the virus. The performance of ARIMA and ARFIMA models in forecasting daily new cases of the disease in Lagos State, Nigeria, was evaluated in this study. The stationarity of the data was tested using the KPSS and ADF tests. To achieve stationarity, the data was subjected to integer and fractional differencing. ARIMA (2,1,1) and ARFIMA (1,0.79,1) were identified using the ACF and PACF plots. The adequacy of the identified models was assessed using the Ljung-Box Chi-Square test. The forecasting performance of both models was compared using Absolute Percentage Squared Error (APSE) and the results show that ARFIMA (1, 0.79, 1) model has a better forecasting performance.

Keywords: COVID-19, ARIMA, ARFIMA, Forecasting performance

1 Introduction

Coronavirus disease (COVID-19) is a contagious disease which causes respiratory disorder which may be mild or severe. The disease was first reported in Wuhan, China (Kandola, 2020). On February 14, 2020, the COVID-19 pandemic was found to have expanded to Africa, with the first confirmed case in Egypt. The first case from Sub-Saharan Africa was reported in Nigeria on February 27th, an imported case from Italy (Lone and Ahmad, 2020). Lagos, Africa's largest city, is the most affected State in Nigeria, recording 58,502 cases. Lagos State is a huge metropolis with a population of almost 20 million people, the majority of whom live in confined quarters where social distance is impossible to maintain. Lagos State is the epic center of COVID-19 in Nigeria and the disease may be prevalent because of the growing population. There is need to understand how the virus tends to grow in the future.

Article History
Received : 28 October 2023; Revised : 30 November 2023; Accepted : 12 December 2023; Published : 30 December 2023

To cite this paper

Oluwagbenga Tobi Babatunde, Chinaza Orji & Abimibola Victoria Oladugba (2023). Modelling Daily New Cases of COVID-19 in Lagos State Nigeria. ARIMA or ARFIMA?. *Journal of Econometrics and Statistics.* 3(2), 185-200.

Time series is a vibrant study subject that has piqued the interest of the academic community in recent decades. The basic goal of time series modeling is to meticulously collect and examine prior data to construct an acceptable model that represents the series' intrinsic structure. The series' future values are then generated using this model (Ratnadip, 2013).

When a time series data is presented for analysis, the general direction of the data is first examined through the plot of the observations. If either upward trend or downward trend is observed, it simply implies non-stationarity of the data. Non-stationarity of the data is caused by variation in the mean and variance (hetereoscedasticity) of the data. In time series analysis, it is appropriate to work with stationary data (Gujarati (2006)). Stationarity here implies that the statistical properties of the data are constant through time (constant mean and variance (homoscedasticity)).

To achieve stationarity, the data must be differenced until it is stationary (usually at most, two differences will be sufficient to make the series stationary and proper care should be taken while differencing to avoid over-differencing.). The variance of the series becomes constant when stationarity is achieved.

In time series modelling and forecasting, there are so many models that are used to model time series data and each of them has its better field of application. The most commonly and frequently used model is the Autoregressive Integrated Moving Average (ARIMA) model. ARIMA model has subclasses; AR (Autoregressive), MA (Moving Average) and ARMA (Autoregressive Moving Average). AR and MA models are the widely used linear time series models in literature. The combination of the two models produces the ARMA model. ARMA model is made up of two components and the values for the components are formally represented in the literature by the letters p and q which is written as ARMA (p, q). The p value refers to the AR component of the model while q is related to the MA component. ARMA is used on a stationary time series data. If the data is non-stationary, stationarity is achieved by taking series of differences on the data; in other words, to make the mean and variance to be constant over time. This demonstrates the ARMA model's shortcoming. Non-stationary time series, which are commonly seen in practice, are insufficiently described by ARMA models. This is the difference between ARMA and ARIMA. The ARIMA model is basically application of ARMA model to non-stationary data. The letter 'I' in the ARIMA model stands for integrated, and it refers to the number of finite differences required to ensure stationarity. If the model does not include any differencing, it is merely an ARMA. An ARIMA process of order (p, d, q) is a model with a d^{th} difference to suit an ARMA (p, q) model. ARIMA's ability to represent a wide range of time series with ease is one of the reasons why it is so widely used.

However, ARIMA is limited to short memory processes. Hence, there exist another model which generalizes ARIMA models. This model is called Autoregressive Fractional Integrated Moving Average (ARFIMA) model. ARFIMA model allows fractional differencing data. It has wide applications especially data with long memory process. Granger and Joyeux (1980) and subsequently Hosking (1981) introduced the use of long memory processes in hydrology, climatology, geophysics, economics, and finance, and it has since been a popular tool for research in these fields. In most cases, ARFIMA outperforms ARIMA in terms of predicting data with long memory process. This is as a result of the differencing parameter in ARFIMA being able to describe both the dependence and short memory structures in data.

Several works in literature have shown that ARFIMA model performs better compared to ARIMA model (see Lucas and Sergio (2016), Yosep and Gumgum (2017), Manohar and Sumalatha (2019)). This study investigates the performance of the ARIMA model and the AFRIMA model in terms of ability to forecast the daily new cases of COVID-19 in Lagos State Nigeria with the purpose of identifying the best model in terms of forecasting ability.

The remainder of the work is arranged as follows: the methodology is discussed in section 2 while the results are presented in section 3. The results are discussed in section 4 while the paper was concluded in section 5.

2 Methodology

The Box and Jenkins (1978) approach introduced by statisticians Box and Jenkins will be employed in fitting the time series models.

2.1 Autoregressive Integrated Moving Average (ARIMA) Model

ARIMA (p,d,q) model is expressed as;

$$\phi(B)(1-B)^d Z_t = \theta(B)a_t \qquad (1)$$

ϕ is the autoregressive parameter

B is the back shift operator

d is the integer differencing parameter

Z_t is the value of the process at time t

θ is the moving average parameter

a_t is the error term or white noise

2.2 Autoregressive Fractional Integrated Moving Average (ARFIMA) Model

The ARFIMA (p,d,q) model is expressed as:

$$\phi(B)(1-B)^d Z_t = \theta(B)a_t \tag{2}$$

ϕ, B, Z_t, θ *and* a_t are as defined (1) above

d is the fractional differencing parameter

2.3 Stationarity/Unit Root Test

The Box and Jenkins approach requires the data under study to be stationary. Two of the common stationarity tests are Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test and the Augmented Dickey Fuller (ADF) test. The test hypothesis for ADF and KPSS tests are given below.

Augmented Dickey Fuller test of hypothesis

H_0: The series is non-stationary

H_1: The series is stationary

Test statistic: $\dfrac{\widehat{\delta}}{se(\widehat{\delta})}$ where se $(\hat{\delta})$ is the standard error for $(\hat{\delta})$.

Decision: Reject the null hypothesis if test statistic > critical value.

Kwiatkowski-Phillips-Schmidt-Shin test of hypothesis

H_0: The series is stationary

H_1: The series is non-stationary

Test statistic: $x_t = r_t + \beta t + e_1$where βt is the trend, r_t random walk and $e_1$is error

Decision: Reject the null hypothesis if test statistic > critical value.

2.4 Estimation of Order of Integration

Several methods of estimating the parameter d of the ARFIMA model exist in literature. However, the approach proposed by Geweke and Porter-Hudak (1983) will be adopted in this study. The bandwidth parameter can be used to change the value of d. The number of bandwidths is traditionally chosen from the interval [$T^{1/2}$, $T^{4/5}$], where T represents bandwidth (Robinson, 1994). The bandwidth of 0.5 is the default and the lowest while bandwidth of 0.8 is the optimal (Hurvich et al. 1998). The default bandwidth of 0.5 will be chosen for this study. The value of d is used to fractionally difference the data once the differenced parameter d has been computed. The remainder of this section is used to describe how to fit an ARMA model to the differenced series.

The ARMA model equation is given as;

$$\phi(B)Z_t = \theta(B)a_t \tag{3}$$

ϕ, B, Z_t, θ *and* a_t are as defined (1) above

2.5 Postulation and Identification of models

The value of the parameter d is determined by the number of differencing required to establish stationarity. Following the guidelines in Table 1 below, the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the stationary series are used to recommend the order of the autoregressive component p and the moving average part q.

Table 1: Behaviour of ACF and PACF.

	AR	MA	ARMA
ACF	Tails off exponentially	Cuts off at lag q	Tails off exponentially
PACF	Cuts off at lag p	Tails off exponentially	Tails off exponentially

Out of the many models that will be identified, the model that obeys the principle of parsimony is considered the appropriate model.

2.6 Estimation of Parameters

The p parameter(s) of the autoregressive ϕ_1, ϕ_2, ---, ϕ_p and q parameter(s) of moving average Θ_1, Θ_2, ---, Θ_q are estimated using Exact Maximum Likelihood (EML) methods with the aid of Gretel software.

2.7 Diagnostic Check

To determine whether the a_t from the model is white noise ($a_t \sim$ IIDN(0, σ^2)), a diagnostic check is necessary. The residuals derived from the models will be analyzed using the Ljung-Box test. The test requirements employ a 0.05 alpha, with a p-value less than alpha indicating that the model is sufficient. The sample ACF and sample PACF of the residual will also be examined to see if the model is adequate. If the ACF and PACF do not have any spikes, we conclude that the model is adequate.

The Ljung Box test statistic is given by;

$$Q = n(n+2)\sum_{j=1}^{m} \frac{r_j^2}{n-j} \qquad (4)$$

Where r_j is the accumulated sample autocorrelations and m is the time lag

2.8 Measuring Forecasting Performance

The forecasting performance of the models will be assessed and compared using the Absolute Percentage Squared Error (APSE) expressed as:

$$\text{APSE} = \sqrt{H^{-1}\sum_{h=1}^{H}(y_{n+h} - \hat{y}_{n+h})^2} \qquad (5)$$

where,

y_{n+h} is the observation at time n+h

$\hat{y}_{n+h}$ is the predicted observation at time n+h n is the number of sample

h is the forecast point

3 Data Analysis and Results

In this study, secondary data on daily new cases of COVID 19 in Lagos State was used. The data was acquired from the Nigeria Centre for Disease Control (NCDC) website (covid19.ncdc.gov.ng and spanned a period of 216 days (March 16, 2020 to October 17, 2020). Figure 1 and Table 2 show the plot and summary of new COVID-19 cases in Lagos State. The plot shows upward and downward movement in the cases of COVID-19 in Lagos State indicating non-stationarity. The results of the ADF and KPSS tests of stationarity are presented in Table 3 and the results of both tests confirmed the non-stationarity of the data. The plot of the differenced series is presented in Figure 2 and it shows that stationarity was achieved after differencing. The ACF and PACF plots of the integer and fractional differenced series are presented in Figures 3 and 4 respectively.

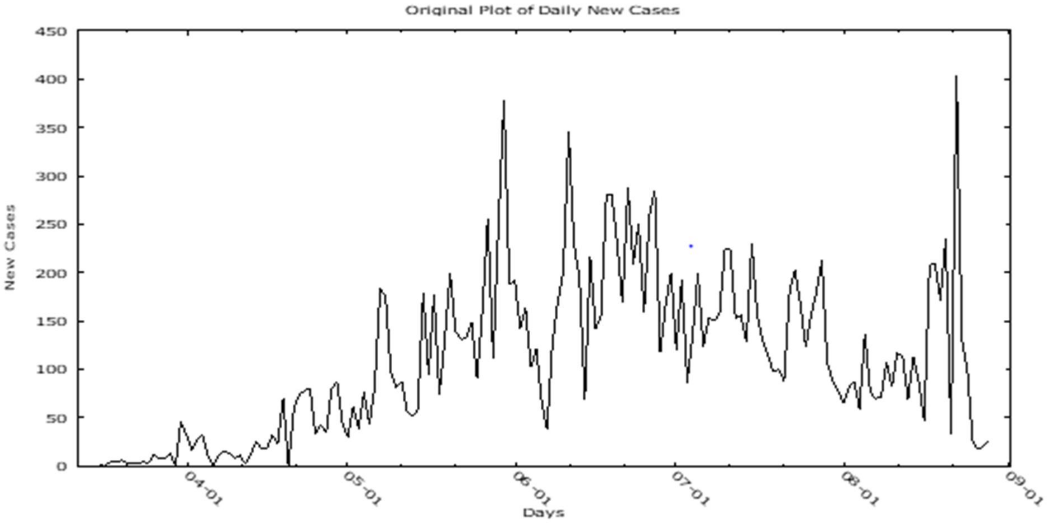

Figure 1: The plot of the daily new cases in Lagos

Table 2: Descriptive Statistics of New Cases of COVID-19 in Lagos State

Mean	95.787
Median	76
Minimum	0
Maximum	404
Standard deviation	79.209

Table 3: ADF and KPSS tests for stationarity

Variable	ADF	KPSS
Test stat	-3.1711	0.5841
Critical value	-2.85 (5%)	0.148 (5%)
Lag	5	4

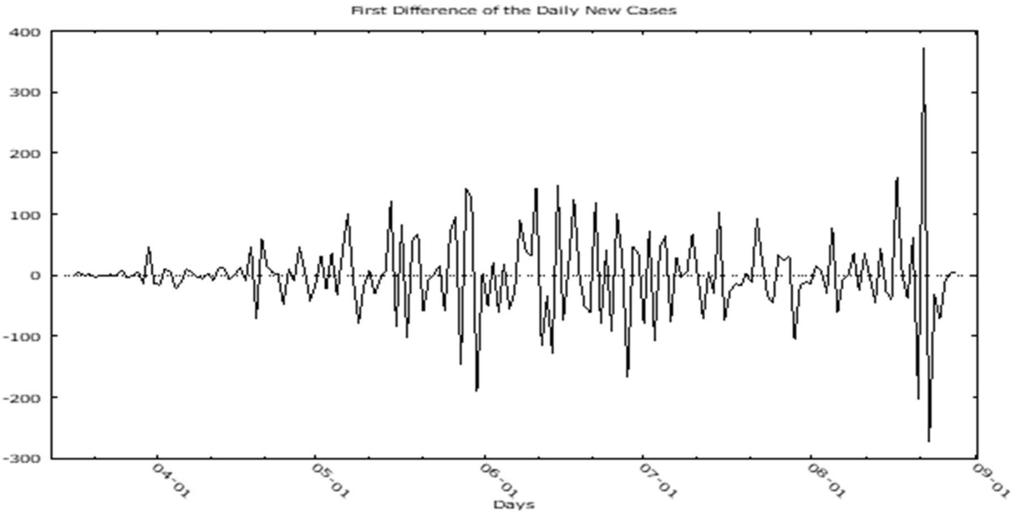

Figure 2: The plot of the first difference

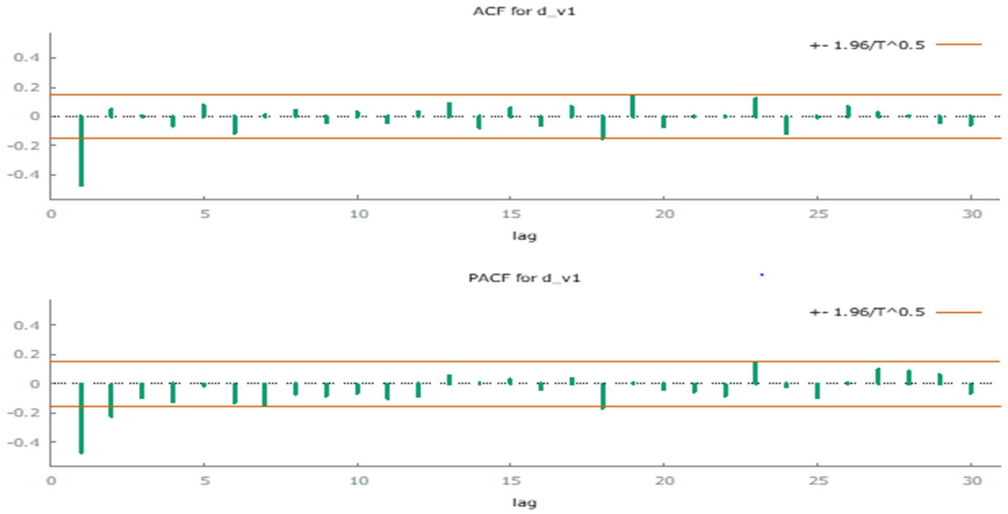

Figure 3: The ACF and PACF plots of the differenced data (d=1)

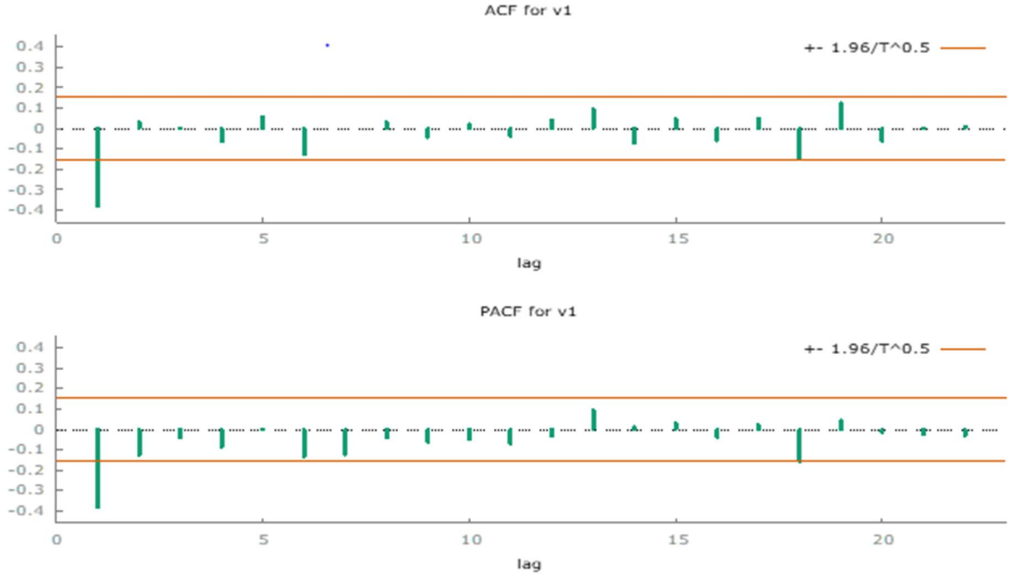

Figure 4: The ACF and PACF of plots of the differenced data (d=0.79)

The ACF and PACF plots in Figure 3 show significant spikes at lag 1 and 2 respectively suggesting ARIMA (2,1,1) while the ACF and PACF plots in Figure 4 show significant spikes at lag 1 and 1 respectively suggesting ARFIMA (1,0.79,1). The parameters of the identified models are presented in Tables 4 and 5.

Table 4: Estimates of the parameters of the ARIMA (2,1,1) model

Parameters	Estimate	p-value
Constant	0.5731	0.4822
Φ_1	0.2251	0.0113
Φ_2	0.1743	0.0410
Θ_1	−0.8963	0.0000

Table 5: Estimates of the parameters of the ARFIMA (1,0.79,1) model

Parameters	Estimate	p-value
Constant	1.9937	0.2497
ϕ_1	0.3299	0.1410
θ_1	−0.7505	0.0000

The parameters of the ARIMA (2,1,1) presented in Table 4 are all significant except the constant term. Therefore, the ARIMA (2,1,1) model is given below.

$$Z_{t_t} = 0.2251Z_{t-1} + 0.1743Z_{t-2} + 0.8963a_{t-1} \qquad (7)$$

The parameters of the ARFIMA (1,0.79,1) presented in Table 5 show that only θ_1 is significant. Therefore, the ARFIMA (1,0.79,1) model is given below.

$$Z_{t_t} = 0.7505a_{t-1} \qquad (8)$$

The adequacy of the ARIMA (2,1,1) model was assessed using the Ljung box test.

H_0: The model is adequate

H_1: The model is not adequate

Testing at alpha = 0.05

Decision rule: reject H_0 if the test statistic > test tabulated

Results: Test statistic (Q) = 8.3912, df = 11, p-value = 0.6779 and test tabulated = 19.675

Conclusion: Since the test statistic is not greater than test tabulated, we do not reject H_0 and conclude that the model is adequate.

The adequacy of the ARFIMA (1,0.79,1) model was assessed using the Ljung box test.

H_0: The model is adequate

H_1: The model is not adequate

Testing at alpha = 0.05

Decision rule: reject H_0 if the test statistic > test tabulated

Results: Test statistic (Q) = 8.8482, df = 11, p-value = 0.6779 and test tabulated = 19.675

Conclusion: Since the test statistic is not greater than test tabulated, we do not reject H_0 and conclude that the model is adequate.

The ACF and PACF plots of the residuals of the ARIMA (2,1,1) model and ARFIMA (1,0.79,1) model are presented in Figure 6 and Figure 7 respectively.

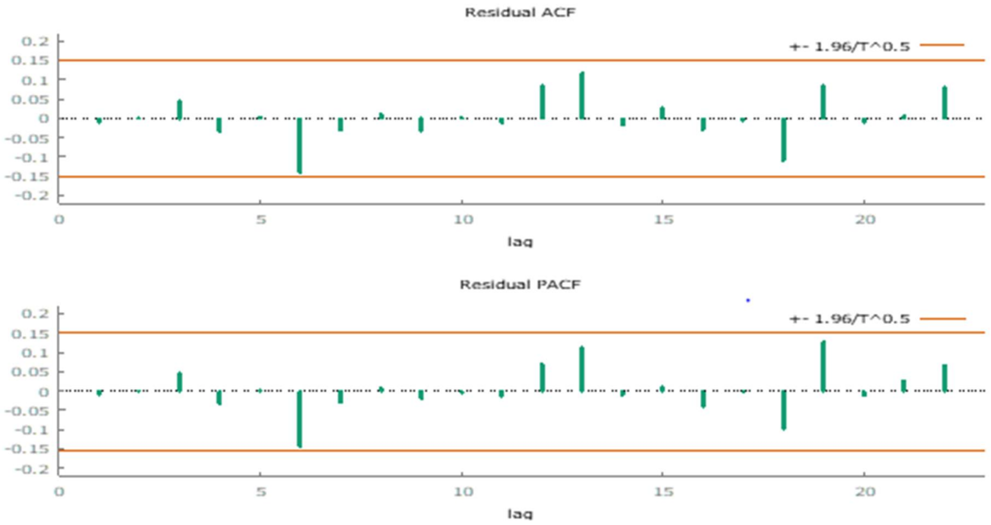

Figure 6: The ACF and PACF plot of the residuals of ARIMA (2,1,1) model

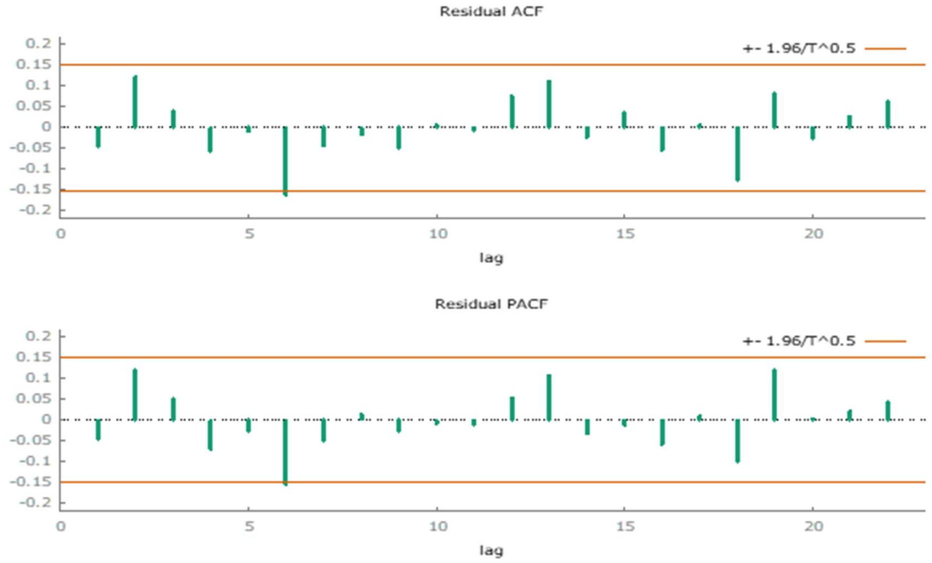

Figure 7: The ACF and PACF plot of the residuals of ARFIMA (1,0.79,1) model

The APSE of both models is presented in Table 6 using (5).

Table 6: Results of the Forecast Performance

Model	APSE
ARIMA (2,1,1)	57.9126
ARFIMA (1,0.79,1)	33.5686

The plots of the out-sample forecast using ARIMA (2, 1, 1) and ARFIMA (1,0.79,1) models are presented in Figure 8 and Figure 9 respectively.

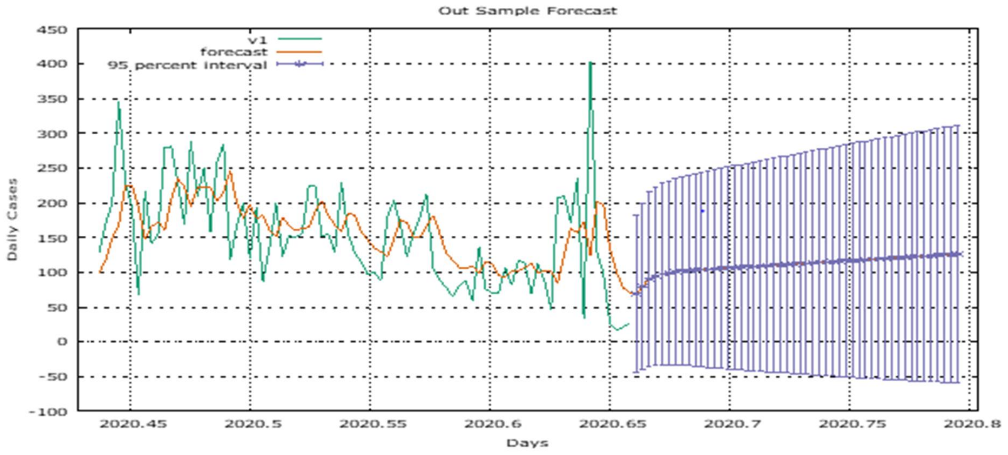

Figure 8: Plot of the out-sample forecast for the ARIMA (2,1,1) model

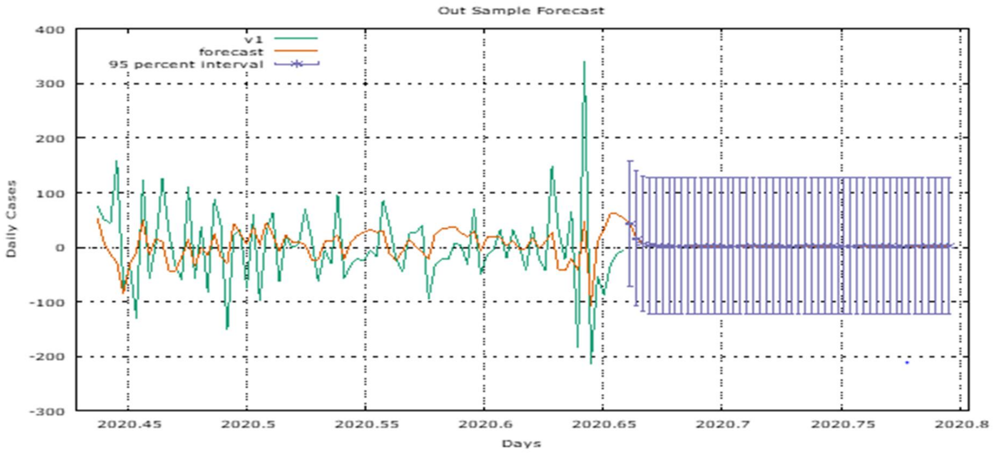

Figure 9: Plot of the out-sample forecast for the ARFIMA (1,0.79,1) model

4 Discussion of Results

This study provided a time series analysis on the daily new cases of COVID-19 in Lagos State. From Figure 1 above, the series is non-stationary because of the upward and downward trend exhibited by the data. It means that the number of cases increased and at some points, it began to decrease slowly and increase again. However, there is need to conduct a formal test to ascertain the stationarity of the data. The stationarity test was performed using the ADF and KPSS tests and the results are presented in Table 3.

Table 3 shows that the data is non-stationary based on the results of both stationarity tests. Integer and fractional differencing were used to difference the data, and the plot of the differenced data is shown in Figure 2. Figure 3 shows the plots of the ACF and PACF for integer differenced data, whereas Figure 4 shows the plots of the ACF and PACF for fractional differenced data. For both integer and fractional differencing, the order of d before stationarity is 1 and 0.79, respectively.

Figure 2 shows that the data points oscillate around a straight line with evidence of no trend indicating that it has a constant mean and a constant variance. This implies that the first differenced data is stationary.

The ACF plot in Figure 3 suggests MA of order 1 while the PACF plot suggests AR of order 2. The auto.arima function in R package was used to select the best model as ARIMA (2,1,1). The parameters of the selected model were estimated and presented in Table 4.

The ACF plot in Figure 4 suggests MA of order 1 while the PACF suggests AR of order 1. The auto.arima function in R package was used to select the best model as ARFIMA (1,0.79,1). The parameters of the selected model were estimated and presented in Table 5.

The adequacy of both models was assessed using Ljung Box test and the results show that both ARIMA (2,1,1) and ARFIMA (1,0.79,1) models were adequate. The adequacy of both models was also assessed through the visualization of the behavior of the residuals of the models. The ACF and PACF plots of the residuals of the ARIMA (2,1,1) and ARFIMA (1,0.79,1) models are presented in Figure 6 and Figure 7 respectively.

The forecasting performance of both models was assessed using the Absolute Percentage Squared Error (APSE) defined in (5) and presented in Table 6. The Absolute Percentage Squared Error (APSE) computed for both models shows that the ARFIMA (1,0.79,1) model has a better forecasting performance compared to the ARIMA (2,1,1) model. Lucas and Sergio (2016) and Manohar and Sumalatha (2019) compared the forecasting performance of ARIMA and ARFIMA models and concluded that the ARFIMA model performed better compared to the ARIMA model.

5 Conclusion

The ARIMA and ARFIMA models were applied to the daily new cases of COVID-19 in Lagos State Nigeria with the purpose of determining which of the models has better forecasting performance. The identified models are ARIMA (2,1,1) and ARFIMA (1,0.79,1). The result of this study shows that ARFIMA (1,0.79,1) model has better forecasting performance compared to the ARIMA (2,1,1) model. This result implies that to have better prediction of the new cases of COVID-19 in Lagos State, ARFIMA model should be employed.

REFERNCES

Box, G. P. F. and Jenkins, G. M. (1978). Time series analysis: Forecasting and Control, 3rd edition, Holden Day, San Fransisco, ISBN-10: 0130607746

Geweke, J. and Porter-Hudak, S. (1983). The estimation and application of long memory time series models, Journal of Time Series Analysis, 4, 221-238.

Granger, C. W. J. and Joyeux, R. (1980). An introduction to long memory time series models and fractional differencing, Journal of Time Series Analysis, 1, 15-39.

Gujarati, D.N. (2006). Essentials of econometrics. New York: McGraw Hill, 4th edition.

Hosking, J. R. M. (1981). Fractional Differencing, Biometrika, 68, 165-176.

Hurvich, C. M., Simonoff, J. S. and Tsai, C. (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion, Journal of the Royal Statistical Society Series B, 60, 271-293.

Kandola, A. (2020, June 30). Coronavirus cause: Origin and how it spreads. Medical News today. https://www.medicalnewstoday.com/articles/coronavirus-causes

Lone, S. A., and Ahmad, A. (2020). COVID-19 pandemic – An African perspective. Emerging Microbes & Infections, 9, 1300–1308.

Lucas, R. T., & Sergio, A. D. (2016). Some comments on fractionally integration processes involving two agricultural commodities. European scientific journal/SPECIAL/edition. ISSN: 1857 – 7881

Manohar, D., & Sumalatha, V. (2019). Time series analysis for long memory process of air traffic using ARFIMA. International Journal of Scientific and Technology Research, 8. ISSN 2277-8616.

Ratnadip, A. (2013). An introductory study on time series modeling and forecasting. Lambert Academic Publishing (LAP) Editor: Dragos. 978-3-659-33508-2.

Robinson, P. M. (1994). Efficient tests of nonstationary hypotheses, Journal of the American Statistical Association, 89, 1420-137.

Yosep, O. S., & Gumgum, D. (2017). The accuracy comparison between ARFIMA and Singular Spectrum Analysis for forecasting the sales volume of motorcycle in Indonesia. AIP Conference Proceedings 1868, 040011 (2017); https://doi.org/10.1063/1.4995126