THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Forecasting COVID-19 Confirmed Cases: Time Series Analysis Perspective

Jubril Oluwatoyin Fantola
Lecturer, Department of Mathematics and Statistics,
The Polytechnic of Ibadan, Ibadan, Nigeria
Olukunmi Olatunji Akanni
Lecturer, Department of Public Health,
Lead City University, Ibadan, Nigeria

Abstract:
COVID-19 is the cause of a new global pandemic threatening millions of lives as it continues to spread internationally. Forecasting the spread of COVID-19 is an analytical and challenging real-world problem to the research community. The study adopted time series model to forecast the cases of COVID-19 using weekly cases of COVID-19 from Nigeria Centre for Disease Control. The model developed and the forecasted results align very closely with the actual number of cases as there is need to put all hands on deck to curtail the spread of the virus.

Keywords: *COVID-19, Pandemic, Cases, Virus, Forecast, Nigeria Centre for Disease Control*

1. Introduction

The recently occurred coronavirus has eliminated masses of people all over the world and is continuously taking people under its arrest. Washing hands, covering your face, isolating hygiene and staying away from the community may be a way to prevent this transmissible disease, but it doesn't seem to be enough to stop the transmission (Nussbaumer-Streit *et al.,* 2020). COVID-19 which stands for "Corona Virus Disease 2019" is a novel highly contagious virus belonging to Coronaviridae family that has been suspected to be transmitted to humans from animals. This virus causes mild to severe respiratory illness and death (Paules *et al.,* 2020). The cases of rapid human-to-human transmission signify that 2019-nCoV is highly infectious than others. Although a local seafood market in Wuhan is believed to be the source of exposure, (Huang *et al.,* 2020). The scope of occurrence of this disease is not clear since its occurrence at present is so dynamic (Paules et al., 2020). An apparent variation is present in epidemiological examinations and detection abilities performed by different countries for detecting infected cases (Niehus *et al.,* 2020). Novel coronavirus is the cause of a new global pandemic threatening millions of lives (Paules *et al.,* 2020). COVID-19 continues to spread internationally. Worldwide, more than 100 000 cases of COVID-19 and more than 3500 deaths have been reported. COVID-19 is thought to have higher death rate than regular influenza, even as wide variation is reported. World Health Organization (WHO) estimates the death rate across the globe at 3.4%, South Korea has noted mortality of about 0.6% (Del Rio and Malani, 2020).

COVID-19 is currently a major international hazard. Forecasting the spread of COVID-19 is an analytical and challenging real-world problem to the research community. Therefore, we use day level information of COVID-19 spread for cumulative cases from Nigeria. The difficulty in modelling and forecasting COVID-19 because it's a real life scenario and has inherent modeling difficulties such as the number of tests, randomness, interventions, stay-at-home compliance, curfews, epidemiological realities, and many other factors contribute to the constraint of forecast models in this case. Countries, especially in Africa who are just witnessing a progressive rise in COVID-19 cases must be decisive in implementing the containment interventions and ensure strict compliance by the citizenry (NCDC, 2020). Forecast are mostly impossible when very little past data exists upon which to conduct such prediction. This study therefore seek to make an attempt in predicting the confirmed cases of COVID-19 based on the already collected weekly data on COVID-19 and the development of an adequate time series model for forecasting the confirmed cases of COVID-19 in Nigeria.

2. Materials and Method

2.1. Dataset Description

In modeling and forecasting the confirmed cases of COVID-19 in this study, a time series data were obtained from the Nigeria Center for Disease Control (http://covid19.ncdc.gov.ng/) from April 1, 2020 to March 30, 2021 on weekly basis.

2.2. Time Series Models

Time series analysis entails methods that breakdown a series into explainable portions that allows trends to be identified, estimated and forecasts to be made. Basically time series analysis attempts to understand the fundamental context of the data points through the use of a model to forecast future values based on known past values. Such time series models include MA, AR, ARIMA, ARMA, GARCH, TARCH, EGARCH, CGARCH and FIGARCH.

2.3. Autoregressive Integrated Moving Average (ARIMA)

In forecasting a time series, ARIMA modeling is one of the best modeling techniques. ARIMA models are always represented with the help of some parameters and the model is expressed as ARIMA (p, d, q). Here, p stands for the order of auto-regression, d signifies the degree of trend difference while q is the order of moving average. ARIMA involves techniques used to analyze timing information in order to obtain important insights and different attributes of information. ARIMA uses models to predict future values that depend on the most recently observed value. ARIMA has typical transient requirements. The ARIMA model is suitable for evidence that the information shows non-stationarity, and the basic difference step can be applied at least multiple times to eliminate non-stationarity (Kwiatkowski *et al.,* 1992).
ARIMA model is given as
$X_t = \varphi + \theta X_{t-1} + \epsilon_t$ ---------------------- (1)

The predicted value X_t depends on the previousprediction $X_{t-1}$and the error ϵ_t calculated as the differencebetween the predicted and actual outcome. $\theta$is the slope coefficient and φ is the nonzero mean.

2.4. Differencing

Differencing simply means subtracting the value of a previous observation from the value of a future observation. Calculating differences among pairs of observations at some lag to make a non- stationary series stationary. There are possible shifts in both the mean and the spread over time for this series. The mean may be edging upwards, and the spread may be increasing. If the mean is changing, the trend is removed by differencing once or twice. If the variability (spread) is changing, the process may be made stationary by logarithmic transformation. The value of *d* is determined by the number of times you have to difference the scores to make the process stationary.
If *d* = 0, the model is already stationary and has no trend.

When the series is differenced once, *d* = 1 and linear trend is removed. When the difference is then differenced, *d* = 2 and both linear and quadratic trend are removed. For non-stationary series, *d* values of 1 or 2 are usually adequate to make the mean stationary.

Box and Jenkins recommend the differencing approach to achieve stationarity. However, fitting a curve and subtracting the fitted values from the original data can also be used in the context of Box-Jenkins models.

3. Results and Observations

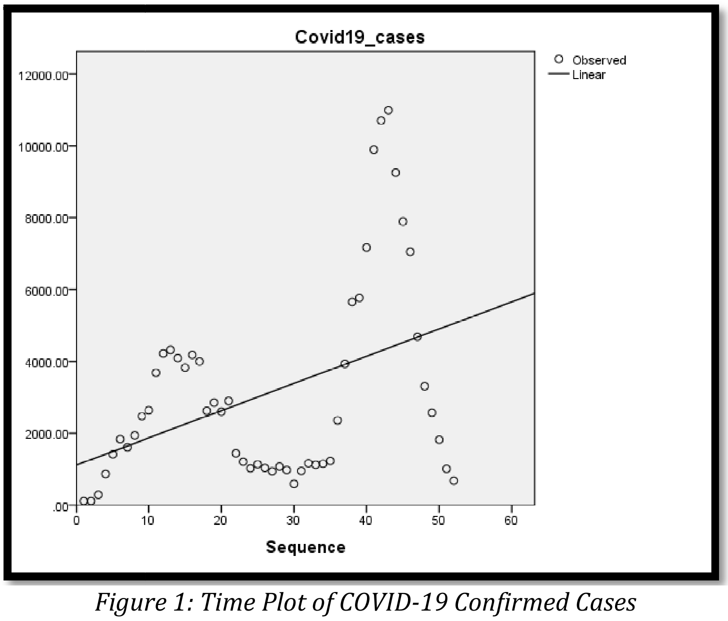

Figure 1: Time Plot of COVID-19 Confirmed Cases

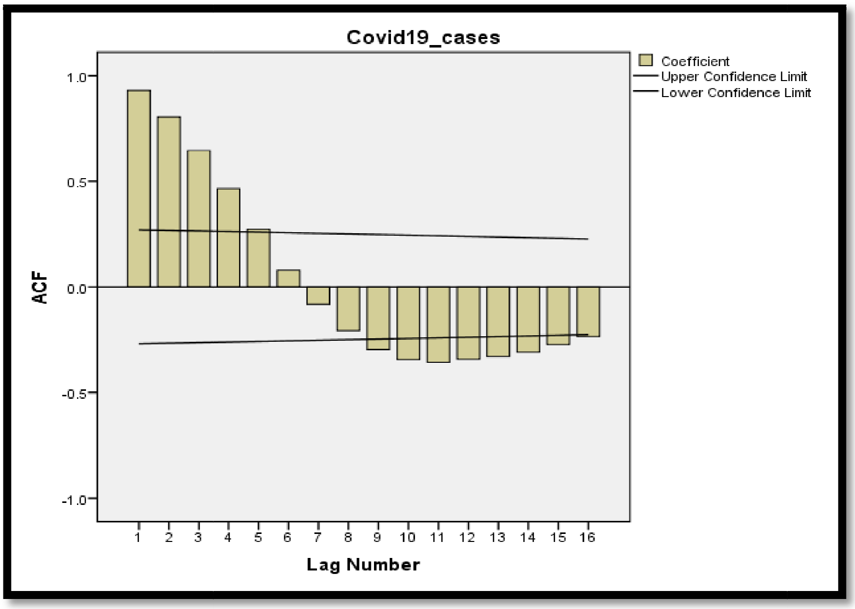

Figure 2: ACF of the Original Series

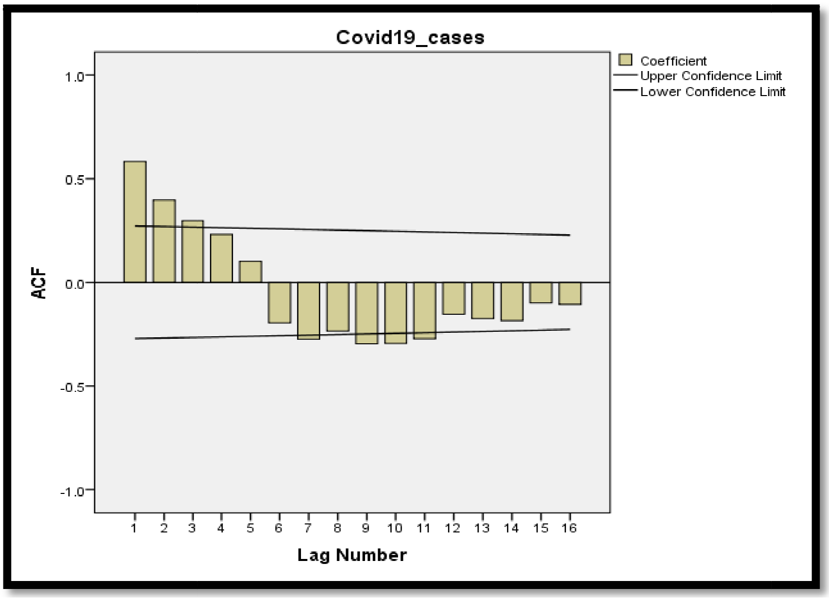

Figure 3: ACF of the First Difference

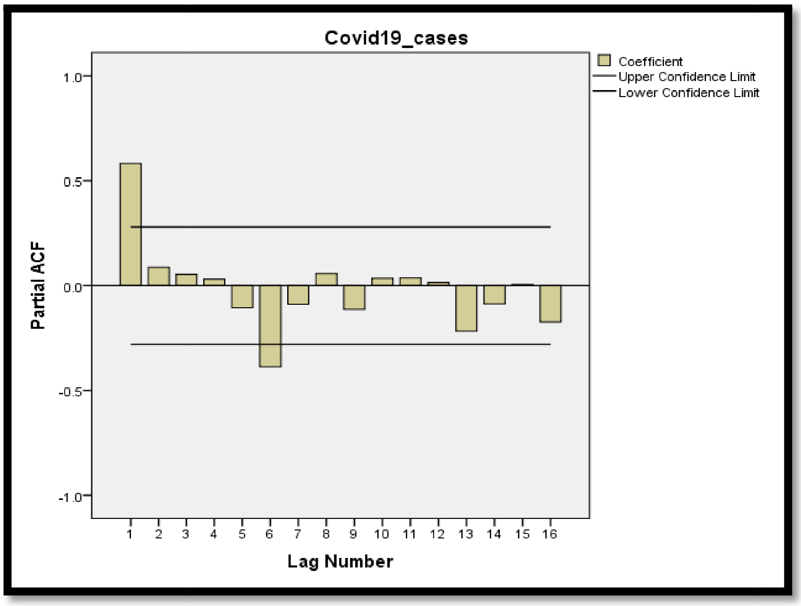

Figure 4: PACF of the First Difference

Model Description	BIC	R Squared
ARI (p,d)		
(1,1)	13.339	0.932
(2,1)	13.428	0.933
(3,1)	13.522	0.933
(4,1)	13.62	0.933
(5,1)	13.71	0.934
IMA (d,q)		
(1,1)	13.462	0.923
(1,2)	13.472	0.93
(1,3)	13.561	0.93
(1,4)	13.648	0.931
(1,5)	13.524	0.945
ARIMA (p,d,q)		
(1,1,1)	13.426	0.933
(1,1,2)	13.524	0.933
(1,1,3)	13.612	0.934
(1,1,4)	13.669	0.936
(1,1,5)	13.614	0.945
(2,1,1)	13.524	0.933
(3,1,1)	13.622	0.933
(4,1,1)	13.719	0.933
(5,1,1)	13.594	0.947

Table 1: Model Identification

Model	Number of Predictors	Model Fit statistics		Ljung-Box Q(18)		
		R-squared	Normalized BIC	Statistics	DF	Sig.
Confirmed_covid19_cases-Model_1	0	.932	13.339	16.286	17	.504

Table 2: Diagnostic Checking

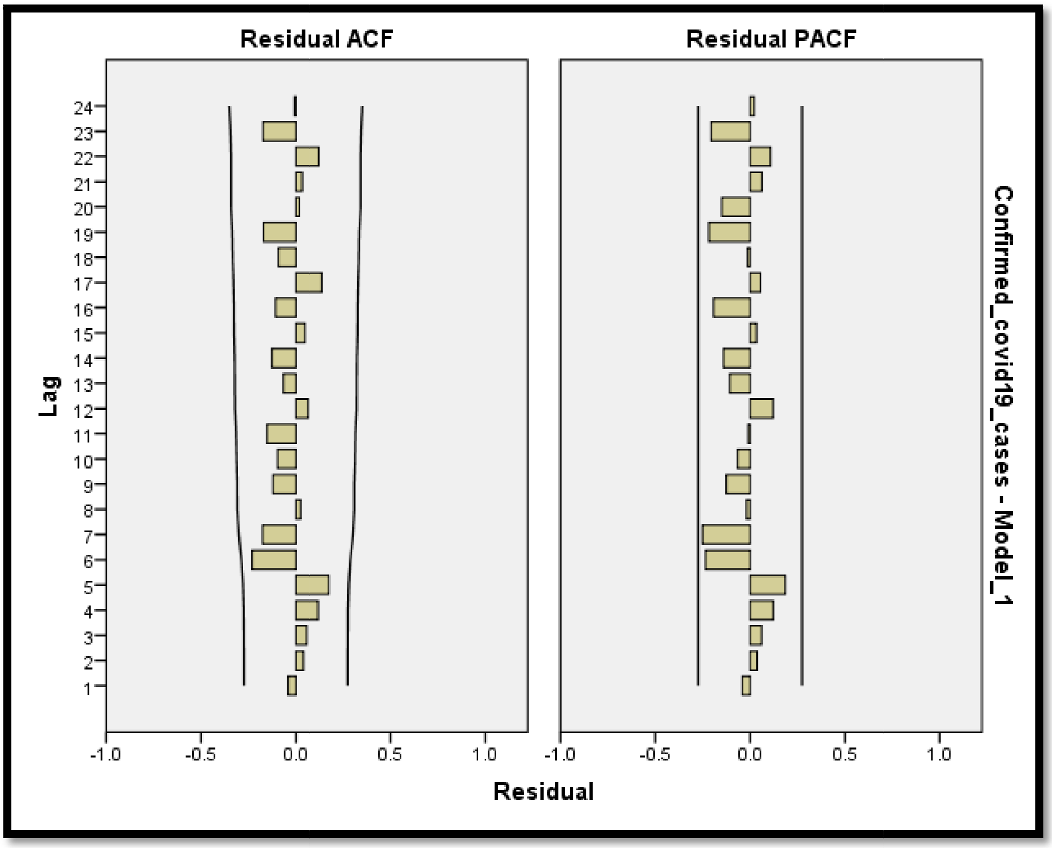

Figure 5: Diagnostic Checking

Weeks	Forecast	UCL	LCL
53	493	1959	-973
54	387	3119	-2345
55	327	4229	-3575
56	294	5254	-4665
57	276	6189	-5637
58	267	7042	-6509
59	262	7826	-7301
60	261	8549	-8028
61	261	9223	-8701
62	262	9853	-9330
63	263	10447	-9920
64	265	11010	-10479
65	267	11545	-11011
66	269	12057	-11519
67	271	12548	-12005
68	274	13020	-12473

Table 3: Model Forecast

The observed series ACF plot shows that there is no seasonality in the data series i.e. data appeared to be exponentially declining. If time series is seasonal the value of kwill not come down to zero quickly as the stationarity will not be achieved easily. The correlogram of the original series using autocorrelation function (ACF), the P-values show that the autocorrelation functions are significant for all lags with P< 0.05, but some autocorrelation functions are high i.e. P> 0.05. We conclude that the series is not stationary. Transforming non-stationary time series to stationary time series at first differencing, the ACF values are low (i.e. P< 0.05)and the Autocorrelation functions are significant for all lags therefore, the series is stationary.

The model identification, the stimulated model ACF and PACF are compared with the original series ACF and PACF to determine the appropriate model and order for the data. The original series of ACF and PACF plot highly resemble the stimulated ARIMA (p,d,f) ACF and PACF plot. This is indicating that the model is Autoregressive and Moving average model. Applying Box Jenkin Methodology, the series was differenced once to make it stationary, i.e. the series integrated is of order 1. Thus, AR1MA (p, d, q) = AR1MA (p, 1, q) model. Since the model is Autoregressive and moving average model, we vary values of p and q then judge the model suitability based on their Bayesian Information Criterion (BIC) and the model with Lowest BIC will be the best model. ARI (1,1) has the lowest Bayesian Information Criterion (BIC) = 13.339. The ARI (1,1) was identified to be the model of the best fit and the model was diagnosed to using the model estimated errors. The diagnostic checks result show that the errors are white noise process with and Box- Ljung test. The P-values of Box-Ljung test is less than 0.05 as the model residuals are independently distribution (white noises) and the model is correctly specified. The model was used to forecast for COVID-I9 cases for sixteen weeks (Table 3)

4. Conclusion

The time plot has the general direction by which a time series appear to be moving over a long period of time. The linear trend using the ordinary least square (OLS) approach predicted an upward movement of the trend. The time series was stationary after differencing. The predicted model showed the number of cases expected in the future weeks as thisinform the Government and stakeholders the need to take appropriate measures in order to reduce the spike on COVID-19.

5. References

i. Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; Cheng, Z.; Yu, T.; Xia, J.; Wei, Y.; Wu, W.; Xie, X.; Yin, W.; Li, H.; Liu, M.; Xiao, Y.; Gao, H.; Guo, L.; Xie, J.; Wang, G.; Jiang, R.; Gao, Z.; Jin, Q.; Wang, J.; Cao, B. (2020). *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, The Lancet , 395, 497.*

ii. Niehus, R.; De Salazar, P. M.; Taylor, A.; Lipsitch, M. (2020). Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travellers, medRxiv.

iii. Nussbaumer-Streit B, Mayr V, Dobrescu AI, Chapman A, Persad E, Klerings I, et al. (2020). Quarantine alone or in combination with other public health measures to control COVID-19: *a rapid review. Cochrane Database Syst Revi.*

iv. Paules C.I, Marston H.D, Fauci A.S. (2020). Coronavirus infections—more than just the common cold. *JAMA. Published. doi:10.1001/jama.2020.0757 ArticlePubMedGoogle Scholar)*

v. Reis, B. Y., Pagano, M., & Mandl, K. D. (2003). Using temporal context to improve biosurveillance. *Proceedings National Academy of Sciences, 100*(4), 1961-1965.

vi. World Health Organization (2020). "Statement on the second meeting of the InternationalHealth Regulations (2005) Emergency Committee regarding the outbreak of novelcoronavirus (2019-nCov),"

vii. Yang Q, Wang J, Ma H, Wang X. (2020). Research on COVID-19 based on ARIMA model-Taking Hubei, China as an example to see the epidemic in Italy. *J Infect Public Health. 13(10):1415–8. https://doi.org/10.1016/j.jiph.2020.06.019*.