EVALUATION OF THE PERFORMANCE OF AUTOREGRESSIVE MOVING AVERAGE (ARIMA)AND ARTIFICIAL NEURAL NETWORKS (ANNS) MODELS USING DAILY CONFIRMED CASES OF COVID-19 IN NIGERIA

Abstract

The COVID-19 outbreak is the most notable world crisis since the Second World War. The pandemic that originated from Wuhan, China in late 2019 affected all the nations of the world and triggered a global economic crisis whose impact will be felt for years to come. This necessitated the need to monitor the deadly disease prevalence for adequate control. The Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) models have been widely used models for forecasting non stationary time series data. These models give good forecasts for future observations but may not be so accurate as expected. This is because the forecasts converge to the mean of the series after multiple forecast values. The artificial neural networks (ANN) can be an alternative method to ARIMA in forecasting the case of nonlinear data like that of daily covid-19 confirmed cases. The forecasting capabilities of ARIMA and ANNs models were compared using the daily COVID-19 confirmed cases in Nigeria. The stationarity of the data was conducted using Augmented Dickey Fuller unit root test while the residual normality test was carried out with the residual plots indicating adequacy of the fitted ARIMA model. The results of neural networks were analyzed using backpropagation for multilayer feedforward powered by sigmoid function. The performance indices for ARIMA and ANNs were compared and the results revealed that the conventional ARIMA model outperformed ANN in terms of minimum prediction error and forecasting ability. The ARIMA (2,1,1) model is affirmed to be best fitted for daily confirmed covid-19 cases in Nigeria over the ANN model.

Keywords: Coronavirus, Performance Evaluation, Stationarity Test, ARIMA Models, Artificial Neural Networks,

1.0 Introduction

Time series analysis in medical and epidemiological area has been conceptual and need special attention in recent times and most especially, during the global pandemic era. Several research studies on case reports of COVID-19 predictions have been conducted with various proposed solution techniques. The prominent techniques fall into two broad categories, namely, statistical and soft computing techniques. Statistical techniques include among others, exponential smoothing, autoregressive integrated moving average (ARIMA), and generalized autoregressive conditional heteroskedasticity (GARCH) model (Wanget al., 2012).

The ARIMA models, also known as the Box-Jenkins models are mathematical models of persistence or autocorrelation in time series. They do not only uncover the hidden patterns in the given data but also generate forecasts and predict a variable's future values from its past values (Akeyede et al., 2022). The use of ARIMA for forecasting time series is essential with uncertainty as it does not assume knowledge of any underlying model or relationships as in some other methods. ARIMA essentially relies on past values of the series as well as previous error terms for forecasting (Tabachnick and Fidell, 2001).

The parameters of ARIMA models are usually estimated by using Ordinary Least Squares (OLS) approach or maximum likelihood method. However, OLS approach imposes strictunderlined assumptions on the model specification in the course of estimating the parametersfor the purpose of achieving significant results. This is traceable to the fact that most data seriesor relations are usually non-linear in the parameter and can also be non-stationary.

An Alternative model that can be used to handle the problem of non-linearity and non-stationarity is Artificial Neural Network (ANN). ANNas a soft computing technique has been used extensively as a forecasting model in many areasof human endeavors such as medical, engineering, economic, business, finance, foreign exchange, and stock problems (Khashei and Bijari, 2009). The neural network is an algorithm that was originally motivated with the intention of

Comment [Ma1]: It is advised for author to include one or two contribution of this study, i.e. which it is concern to help particular institution, for example: "the prediction of time series data in medical and epidemiological could help government of Nigeria in manage the out-break of dangerous virus in future"

Comment [Ma2]: What techniques for soft computing?

My suggestion, firstly, author should tell one or two techniques, or directly state ANN as one of soft computing technique. Then, author may continue with "Statistical techniques include among others,..."

having machines that can mimic the brain. An ANN consists of an interconnected group of artificial neurons that are physical cellular systems capable of obtaining, storing information and using experiential knowledge.

In human brain,neuron system and learning process include the modifications to the synaptic connections between the neuron. In the same manner, an ANNs adjust their structure based on output and input information that flows through the network during the learning phase. Similarly, in a typical data processing procedure, ANN has two main steps which are learning and application phases. The models have the capability to form complex nonlinear systems needed for forecasting based on sampled data. ANNs are self adaptive and data driven models that require little assumptions, but are capable of learning from the data and make a generalization from about the observations from the results obtained. The models are also referred to as "universalapproximator" with ability to estimate any continuous mapping to the preferred degree of precision (Shamsuddeen and Muhammad, 2020).
An artificial neural network is a massively parallel combination of simple processing units which can acquire knowledge from an environment through a learning process and store the knowledge in its connections

This study therefore intends to contribute to a better understanding of the problem related to the covid-19 daily cases prediction. In order to achievethis, linear models (ARIMA) is compared with non-linear models (ANNs) to establish if the forecast power improves from one technique to other. Thus, the aim of this study is to model and predict the daily confirmed cases of covid-19 data in Nigeria using the linear (ARIMA) and nonlinear (ANNs) models.

Comment [Ma3]: Please recheck whether ARIMA is a linear model or not. As for my knowledge, we do not called ARIMA as linear model because it is constructed with linear component and non-linear component. So, please do recheck whether it is suitable term to put "linear" here.

Additional, we knew that in real world, the sample data mostly has non-linear patterns, so, my suggestion, author should elaborate the idea based on non-linear data clearly and theoritacally.

2.0 Related Literature Review

2.1 Background to COVID-19 Cases

In December 2019, a novel Corona-virus emerged at a livestock Market in the Wuhan State of Hubei Province in China and has evolved into a global. The Chinese authorities announced the isolation of this novel virus on January 7, 2020 and it was named COVID-19 by the World Health Organization (WHO) on January 12, 2020. As at February 12, 2020, a total of 43,103 confirmed cases and 1,018 deaths were been reported (WHO, 2020).

From January 16 to February 14, 2020, the cumulative number of new confirmed cases of COVID-19 in mainland China was 50 031, including 37 930 in Hubei province, 22 883 in Wuhan city and 12,101 in other provinces outside Hubei. The peak of the number of new confirmed cases in other provinces outside Hubei was from January 31 to February 4, 2020, and the peak of new confirmed cases in Wuhan city and Hubei province was from February 5 to February 9, 2020. The number of new confirmed cases in other provinces outside Hubei showed a significant decline (23% compared with the peak) from February 5 to February 9, 2020, while the number of new confirmed cases in Wuhan city (30% compared with the peak) and Hubei Province (37% compared with the peak) decreased significantly from February 10 to February 14, 2020.

Fanelli and Piazza, (2020) analyzed the temporal dynamics of COVID-19 outbreak in China, Italy and France with the time-frame of January 22 to March 15 2020. The first analysis of simple day-lag maps points to some universality in the epidemic spreading. The analysis of the same data within a simple susceptible-infected-recovered-deaths modelindicated that the kinetic parameter that described the rate of recovery. This appeared to be the same regardless of the country where the infection and death rates appeared to be more prevalent.

The model placed the peak in Italy around March 21, 2020, with a peak number of infected individuals of about 26,000 (not including recovered and dead) and a number of deaths at the end of the epidemics of about 18,000. Since the confirmed cases were believed to be between 10% and 20% of the real number of individuals who eventually get infected, the apparent mortality rate of COVID-19 fell between 4% and 8% in Italy, while it appeared substantially lower, between 1% and 3% in China.

Piccolomini and Zama (2020) also proposed the modification of the Susceptible-Exposed-Infected-Recovered-Dead (SEIRD) differential model for the analysis and forecast of the COVID-19 spread in some regions of Italy. They introduced a time-dependent transmitting rate and reported the maximum infection spread for the three Italian regions firstly affected by the COVID-19 outbreak.

On February 27 2020, Nigeria recorded its first case of Covid-19. The index case was an Italian citizen who arrived Nigeria through the Murtala Muhammed International Airport, Lagos aboard Turkish airline from Milan, Italy. On March 9 2020, the second case of Covid-19 was reported in Nigeria as contact of the index case, and as the days and weeks progressed, the number of confirmed cases of Covid-19 increased stemming from both local transmission and importation from other countries.

On the mode of transmission of COVID 19, it is contacted through droplets which remain in the air for some period of time, which also occurs through human interactions and contaminated fomites. The inhabitants from countries like United Kingdom, Germany, Italy Spain, USA, etc were considered as people with high risk factors (Nwafor et al., 2022; Pullano et al., 2020).

For the treatment of COVID 19, there exists not any available therapeutic product which is effective for the cure of the disease. The virus could only be managed by a number of discovered medicines (Adhikari and Meng, 2020).
As of March 29 2020, the total confirmed cases within Nigeria had risen to 97 (ninety-seven) and had recorded its first death on March 23 2020.

The Federal Government on March 29, 2020, announced a lockdown in Lagos, Ogun states and the Federal Capital Territory (FCT) with effect from 11pm of March 30 2020. Lagos State was the epicentre of the disease and Ogun state being its boarder state while FCT had the second-highest of confirmed Covid-19 cases in the country. The government had also imposed travel restriction into the country for travellers from China, Italy, Iran, South Korea, Spain, Japan, France, Germany, the US, Norway, the UK, Switzerland and the Netherlands on the March 8. It expanded these restrictions on March 21 as the nation closed its two main international airports in Lagos and Abuja. The country also suspended rail services on March 23, 2020.

As the number of cases grew nationwide in Nigeria and local transmission surged relative to the number of imported cases, there was the need to focus on more local measures to decrease the spread of COVID-19. In addition to the lockdown, social distancing rule was enforced by cancelling mass gatherings, closing businesses except for providers of essential goods and services.

2.2 On ARIMA Models

The time series ARMA model was used by Rediat (2020) to model projection of COVID 19 prevalence cases in East African countries such as Sudan, Ethiopia, Djibouti and Somalia. The research results indicated that progression in the frequency of cases in area of study
Fang and Berlin (2014) employed transfer function model to analyze and forecast students' study achievement, and the result indicated that ARIMA model showed stability and accuracy, transfer function showed better accuracy to predict students' academic performance in college examinations.

Ceylan (2020) modeled and forecasted the epidemiological trend of COVID 19 occurrence in European critically affected countries with ARIMA having the lowest mean absolute percentage error values.

Applying ARIMA model in Turkey on the day 150 of COVID 19 disease, Unvan and Demirel (2020) forecasted that the number of confirmed cases will not cut to zero level until 6 August 2021.

Gupta and Pal (2020) fitted ARIMA model to predict the number of confirmed cases in India in both worst and optimistic scenarios. The worst cases scenario was predicted to have nearly about 700,00 cases by the end of by April, 2020.

2.3 On Artificial Neural Networks

Shamsuddeen and Muhammad (2020) compared the prediction ability of ARIMA and ANN models, and the empirical results indicated that the artificial neural network model gives better predictions and forecasts over the ARIMA model.

Nawaf et al. (2021) proposed an artificial neural network model to estimate and forecast the number of confirmed and recovered cases of COVID 19. The proposed model was based on the training data published in Saudi Arabia COVID-

19 demography using multilayer perception neural network. The results revealed that the number of recoveries could be between 2000 and 4000 per day.

Dilbag et al. (2020) considered classification of COVID-9 patients from chest CT images using multi-objective differential evolution-based convolution neural networks. An extensive analysis showed that the model could classify the chest CT images at a good accuracy rate.

Liao and Wang (2010) used a stochastic time effective neural networks in predicting China global index and their study showed that the mentioned model outperformed the regression model.

Hyup Roh (2007) introduced hybrid models with neutral networks and time series model for forecasting the volatility of stock price index in two vision points: deviation and direction and the results showed that ANN time series models can increase the predictive power for the perspective of deviation and direction accuracy.

3.0 Methodology

3.1 Data Source

The data used for this study were obtained from publications of the World Health Organization (https://ourworldindata.org/covid-cases) which were daily updated cases of COVID-19 around the world. The observations were daily time series of Nigeria's daily confirmed cases of Covid-19 from February 28, 2020 to November 30, 2021

3.2 Method of Data Analysis

3.2.1 Stationarity Test

Before applying any statistical model it is important to check if our data is considered as stationary. Stationarity basically means that the properties such as the mean and variance don't change over time. In this study, we used the Test Dickey-Fuller, which states that if the p-value is lower than a given threshold (level of significance) it will not be considered as stationary.

ARIMA and Artificial Neural Networks are the two forecasting techniques deployed in this study to predict the daily confirmed cases of covid-19 data in Nigeria. The R software is also used for the data analysis.

3.2.2ARIMA model

ARIMA model uses current values for the forecasting of the future values in the series such that the following assumptions hold:(i) Data stationarity (ii) univariate data requirement(Ajao et al., 2020).

The ARIMA (p, d, q) is represented by the form

$$\theta_p(A)(1-A)^d z_t = \emptyset_q(A)b_t, \quad (1)$$

where p is the Autoregressive (AR) lag order, d is the differencing order and q is the Moving Average (MA) lag order.

Thus,

$$\theta_p(A) = (1-\theta_1 A-\theta_2 A^2-\cdots-\theta_p A^p) \quad (2)$$

$$\emptyset_p(A) = (1-\emptyset_1 A-\emptyset A^2-\cdots-\emptyset A^p) \quad (3)$$

The seasonal pattern of the general form of ARIMA is given as

$$\theta_p(A)\theta_p(A^s)(1-A)^d(1-A^s)^d z_t = \emptyset_q(A)\theta\emptyset(A^s)_{b_t}, \quad (4)$$

where s is the seasonal period

Equations (2) and (3) can now be respectively re-written as

$$\theta_p(A^s) = (1-\theta_1 A^s-\theta_2 A^{2s}-\cdots-\theta_p A^{ps}) \quad (5)$$

$$\theta_q(A^s) = (1-\theta_1 A^s-\theta_2 A^{2s}-\cdots-\theta_q A^{qs}) \quad (6)$$

3.2.3ARIMA Model Building

Comment [Ma4]: Parameter p and q need to be written in italic form

Comment [Ma5]: "ARIMA Model Building" is not suitable term. Because ARIMA is already designed and build. My suggestion:
"ARIMA Modeling Procedure"

The Box and Jenkins' modeling procedure consists of three iterative stages which are specification of the model, parameter estimation and diagnostic checking. The process is repeated many times up to when an acceptable model is obtained. The selected values can then be used to predict the future value of the data (Box et al., 1994).

(a) Specification of the model: This stage ensures that the time series variables are made stationary through the process of differencing. The graphs of autocorrelation function and that of partial autocorrelation functions are plotted to decide AR and MA components for further analysis. For a pure autoregressive process of lag p, the partial autocorrelation functions up to lag p will be the autoregressive coefficients while beyond that lag; we expect them all to be zero. In general, there will be a 'cut of' at lag p in the partial autocorrelation function.
The correlogram on the other hand will decline asymptotically towards zero and not exhibit any discreet 'cut of' point. An MA process of order q, on the other hand, will exhibit the reverse property.

(b) Parameter estimation: This is simply the process of estimating the parameters of the model using adopted computational algorithm. The maximum likelihood method of estimation is often used for the estimation of the model's parameters.

(c) Diagnostic checking:This is the process of testing whether or not the estimated model adequately specifies a stationary univariate process. It is adjudged that an estimated model is refined if the fitted model is adequate.

3.3.4Methods of Model Identification

Akaike Information Criterion (AIC): This is a single member score that can be used to determine which of the multiple models is most likely to be the best model.

$$AIC(p) = nln(\frac{\hat{\sigma}_\epsilon^2}{n}) + 2p \quad (7)$$

Bayesian Information Criterion (BIC): This measures the efficiency of parametized model in terms of predicting the data.

$$BIC(p) = nln\left(\frac{\hat{\sigma}_\epsilon^2}{n}\right) + p + pln(n)\,, \quad (8)$$

where n is the number of observations sampled for model fitting, p is the total number of parameters used by the model is the sum of sample variance. The Smaller the values of AIC and BIC, the better the model

3.3.5Evaluation of Forecasting Methods

The forecasting results of the models are evaluated using Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The model with the least error is chosen as the best:'

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \quad (9)$$

$$RMSE = (n^{-1}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2)^{\frac{1}{2}} \quad (10)$$

$$MAE = n^{-1}\sum_{t=1}^{n}|y_t - \hat{y}_t| \quad (11)$$

3.3.6Artificial Neural Networks - Modeling Procedure and Algorithm

The neural networks used in this study are feed-forward multi-layer perceptions which employ a sigmoid transfer function. The results of four feed-neural networks are analyzed as follows:

1. Quick method: When the quick method is selected, a single neural network is trained. By default, the network has one hidden layer containing max $[3(n_i + n_o)] / 20$ neurons, where n_i is the number of input neurons and n_o is the number of output neurons;
2. Dynamic method: The topology of the network changes during training, with neurons added to improve performance until the network achieves the desired accuracy. There are two stages to dynamic training: finding the topology and training the final network;
3. Multiple method: Multiple networks are trained in pseudo parallel fashion. Each specified network is initialized, and all networks are trained. When the stopping criterion is met for all networks, the network with the highest accuracy is returned as the final model. That is, at the end of the training, the model with the lowest RMS error is presented as the final model;

4. Prune method: This is conceptually the opposite of the dynamic method. Rather than startingwith a small network and building it up, the prune method starts with a large network and gradually prunes it by removing unhelpful

neurons from the input and hidden layers. Pruning proceeds in two stages: pruning the hidden neurons and pruning the input neurons.

3.3.7 The Backpropagation Learning Algorithm of Neural Network

The backpropagation procedure uses the gradient descent learning technique for multilayer feedforward. The training network consists of two main parts, the input and output parts. Initial weights are selected randomly between –1 and +1. The network outputs depend on the input units, hidden units, weights of the network, and the activation function. The difference between the computed and the actual output (target) is known as network error. Back propagation method takes the network error and propagates it backward into the network. Errors are used at each neuron to update weights. This process is repeated until the total network error becomes the smallest.

ANN method uses the error function to measure the difference between the target value and the output value. The weights of the network are frequently adjusted in such a way that the error function becomes as small as possible. The target function can be written as $E_p = T_p - Y_p$, where T_p is the targeted output value of the p[th] iteration, and Y_p is the computed output of the p[th] iteration

3.3.8 Backpropagation for Multilayer Feedforward

Backpropagation method is based on three factors: the learning rate, the distance between the actual output and predicted output, and the activation function. The learning rate controls the size of change in weights in each step. If it is too small, the ideal point of convergence may be small. But in the case if the learning rate is too large, the algorithm might not converge at all. The learning rate should fall in the range $0 \leq a \leq 1$.

Let the error function be denoted by E_f and the rate of change in E_f with respect to the weight, β be written as

$$\Delta E_f(\beta) = \frac{\partial E_f}{\delta \beta_p}, \quad (12)$$

where β_p is the vector of all weights of the network at p[th] iteration.

The network weights are determined by

$$\beta_{p+1} = \beta_p + \Delta(\beta)_p, \quad (13)$$

where β_p are weights of the p[th] iteration, β_{p+1} are the parameters of $(p+1)^{th}$ iteration and $\Delta(\beta)_p$ is the learning process which can be expressed as

$$\Delta(\beta)_p = -a\nabla E_f(\beta), \quad (14)$$

where a referred to as learning rate is a positive constant.

Let $E_f = \frac{1}{2}\sum_{p=1}^{n}(T_p - Y_p)^2$ be a network objective function so that

$$Y_p = f(\sum_{j=1}^{m}\sum_{i=1}^{n} X_i v_{ji}) = f(X_p \beta_p) \quad (15)$$

From equation 15, the previous equation (12) can be rewritten as

$$\Delta E_f(\beta) = \frac{\partial E_f}{\partial \beta_p} = \frac{\partial E_f}{\partial v_{ji}} \quad (16)$$

Using sigmoid function as the activation function in ouput layer, we generate

$$f(u) = \frac{1}{1+e^{-u}}, \quad (17)$$

where

$$u = \sum_{j=1}^{m}\sum_{i=1}^{n} X_i v_{ji} \quad (18)$$

The gradient $\frac{\partial E_f}{\partial v_{ji}}$ can be expressed as

$$\frac{\partial E_f}{\partial v_{ji}} = \frac{\partial E_f}{\partial f(u)} \cdot \frac{\partial E_f}{\partial (u)} \cdot \frac{\partial E_f}{\partial v_{ji}} = f(u)(1 - f_p(u)(T_p - f_p) \quad (19)$$

From equation (18), we derive

$\frac{\partial u}{\partial v_{ji}} = X_i(20)$

In order to analyze the residual, the value of $\frac{\partial E_f}{\partial f(u)}\frac{\partial f(u)}{\partial (u)}$needs to be determined as follows:

$$\frac{\partial E_f}{\partial f(u)} = \frac{\frac{\partial}{2}(T_p - f_p(u))^2}{\partial f(u)} = \left(T_p - f_p(u)\right)\frac{\frac{\partial}{2}(T_p - f_p(u))}{\partial f(u)} = -(T_p - f_p(u)) \quad (21)$$

Therefore,

$$\frac{\partial f(u)}{\partial u} = -f_p(u)\left(1_p - f_p(u)\right)(22)$$

Substituting equations (20), (21) and (22) into equation (19), we have

$$\frac{\partial E_f}{\partial v_{ji}} = \frac{\partial E_f}{\partial f(u)} = \frac{\partial f(u)}{\partial (u)}.\frac{\partial (u)}{\partial v_{ji}} = -f_p\left(1 - f_p(u)\right)(T_p - f_p(u))X_i(23)$$

Equation (19) is rewritten as:

$$\Delta E_f(\beta) = \frac{\partial E_f}{\partial \beta_p} = \frac{\partial E_f}{\partial v_{ji}} = -X_i f_p(u)(1 - f_p(u)(T_p - f_p(u)) \quad (24)$$

Substituting (24) into(16) and subsequently into (15), we generate the frequency of the iterations and $\Delta(\beta)_p$of the learning process as

$$\beta_{p+1} = \beta_p + aX_i f_p(u)\left(1 - f_p(u)\right)\left(T_p - f_p(u)\right)$$

$$= \beta_p + aX_i f_p(u)(X_p, \beta_p)\left(1 - f_p(X_p, \beta_p)\right)\left(T_p - f_p(X_p, \beta_p)\right), \quad (25)$$

where a is the learning rate, $\left(T_p - f_p(X_p, \beta_p)\right)$ is the distance between the actual output and the predicted output and $f_p(u)(X_p, \beta_p)$ is the activation function.

3.3.9 Comparison of ARIMA with ANNs Results

To compare the results of the two models under consideration, we follow the approach of analyzing the mean errors in the models. The minimum error and the maximum error between the observed values and the estimated values are also considered. Values close to +1.0 indicate a strong positive association, so that high predicted values are associated with high actual values and low predicted values are associated with low actual values. Values close to –1.0 indicate a strong negative association, so that high predicted values are associated with low actual values, and vice versa. Values close to0.0 indicate a weak association, so that predicted values are more or less independent of actual values.

Comment [Ma6]: Please follow the format of writing.

4.0 Results and Discussion

4.1 Descriptive Statistics

The total number of observations in terms of confirmed cases of Covid-19 from February 28, 2020 to November 30, 2020 was n = 642 days.

Table 1: Descriptive Statistics of COVID-19 Cases in Nigeria

			Statistic	Std. Error
new_cases	Mean		333.6729	14.99588
	95% Confidence Interval for Mean	Lower Bound	304.2259	
		Upper Bound	363.1199	
	5% Trimmed Mean		284.7087	
	Median		188.5000	
	Variance		144370.717	
	Std. Deviation		379.96147	
	Minimum		.00	

	Maximum	2464.00	
	Range	2464.00	
	Interquartile Range	411.00	
	Skewness	2.003	.096
	Kurtosis	4.672	.193

From the descriptive statistics, the mean of the time series equal 333.6729;the median of the time series is 188.5000;the standard. error of the time series is 379.96147 and the sum of the observations is 642. However, the standard error is observed to be higher than the mean, suggesting a very high volatility in the series.

Comment [Ma7]: standard error

Table 2: Outlier Test

Extreme Values

			Case Number	Value
new_cases	Highest	1	331	2464.00
		2	329	1964.00
		3	338	1883.00
		4	323	1867.00
		5	335	1861.00
	Lowest	1	638	.00
		2	636	.00
		3	634	.00
		4	628	.00
		5	595	.00[a]

As observed from the table of extreme values above, the series have at least 5 highest and lowest values which implies that presence of outliers in series. This is however not good for the application of the modeling techniques and thus may lead to non-robust modeling performance.

Table 3:Tests of Normality of Series

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
new_cases	.190	642	.000	.782	642	.000

The normality testsare significant since their respective p-values are less than the level of significance showing that the series is not normally distributed at 5% level of significant. The series may also not be stationary and therefore will have to be transformed logarithmically to correct this defect.

4.2 Series Exploration

To further understand the behavior of the series and possible trend, the datasets are plotted over time.

Figure 1: Series Plot on Daily Basis

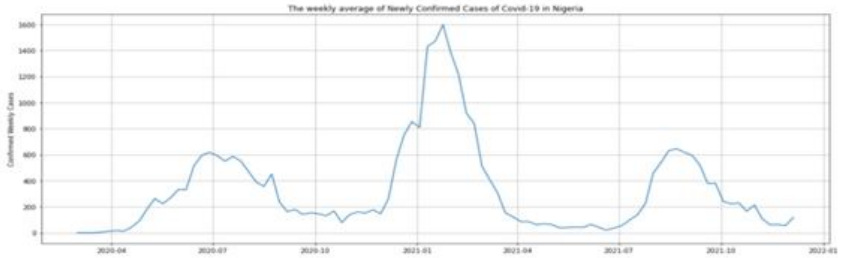


Figure 2: Series Plot of Weekly Basis

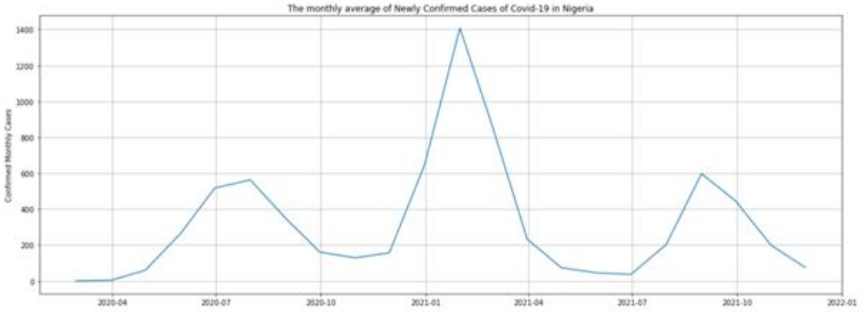


Figure 3:Series Plot on Monthly Basis

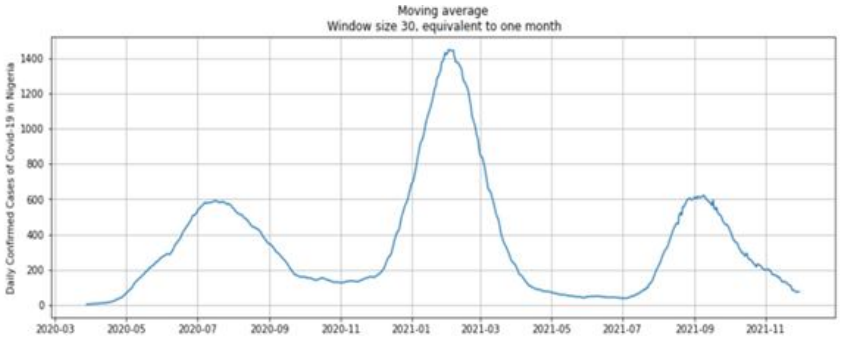


Figure 4: Series Plot of Monthly Moving Average

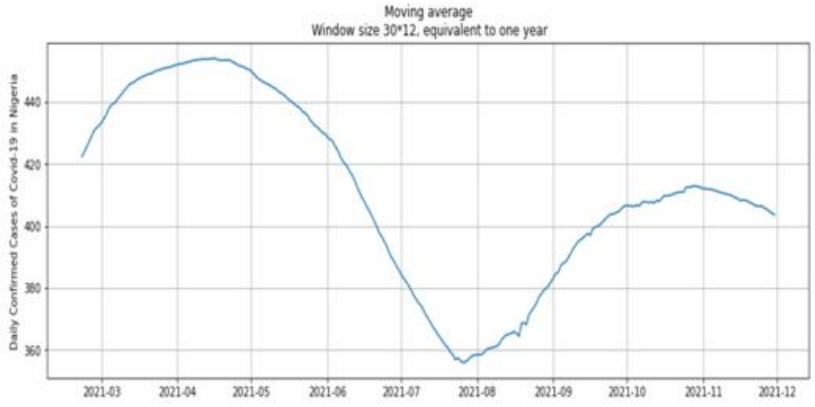


Figure5:Series Plot of Yearly Moving Average

4.3 Stationarity Test Results

The Dickey-Fuller test was conductedand the result presented in Table 4 below

Table 4 : Dickey-Fuller Test at ∝ Level

Results of Dickey-Fuller Test:	
Test Statistic	-2.955419
p-value	0.039273
#Lags Used	18.000000
Number of Observations Used	623.000000
Critical Value (1%)	-3.440890
Critical Value (5%)	-2.866190
Critical Value (10%)	-2.569247
dtype: float64	

Data is non-stationary

Results of Dickey-Fuller Test:	
Test Statistic	-4.816889
p-value	0.000050
#Lags Used	15.000000
Number of Observations Used	626.000000
Critical Value (1%)	-3.440839
Critical Value (5%)	-2.866168
Critical Value (10%)	-2.569235
dtype: float64	

Data is stationary

Table 5: ARIMA Candidate Models

```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=8439.895, Time=0.95 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=8761.770, Time=0.03 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=8610.302, Time=0.06 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=8448.350, Time=0.26 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=8759.772, Time=0.01 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=8451.611, Time=0.38 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=8438.181, Time=0.48 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=8448.454, Time=0.26 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=8496.888, Time=0.08 sec
 ARIMA(3,1,1)(0,0,0)[0] intercept   : AIC=8440.007, Time=0.64 sec
 ARIMA(3,1,0)(0,0,0)[0] intercept   : AIC=8473.101, Time=0.10 sec
 ARIMA(3,1,2)(0,0,0)[0] intercept   : AIC=8441.895, Time=0.94 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=8436.187, Time=0.21 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=8446.460, Time=0.12 sec
 ARIMA(2,1,0)(0,0,0)[0]             : AIC=8494.890, Time=0.04 sec
 ARIMA(3,1,1)(0,0,0)[0]             : AIC=8438.012, Time=0.27 sec
 ARIMA(2,1,2)(0,0,0)[0]             : AIC=8437.963, Time=0.30 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=8608.303, Time=0.03 sec
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=8449.614, Time=0.24 sec
 ARIMA(3,1,0)(0,0,0)[0]             : AIC=8471.103, Time=0.05 sec
 ARIMA(3,1,2)(0,0,0)[0]             : AIC=8439.900, Time=0.39 sec

Best model:  ARIMA(2,1,1)(0,0,0)[0]
Total fit time: 5.876 seconds
```

The Model with the least AIC is considered the best ARIMA model to fit the dataset, which is thus observed to be ARIMA (2,1,1).

Table 6:Summary of ARIMA Model Results

ARIMA Model Results

Dep. Variable:	D.y	No. Observations:	611
Model:	ARIMA(2, 1, 1)	Log Likelihood	742.058
Method:	css-mle	S.D. of innovations	0.072
Date:	Wed, 01 Dec 2021	AIC	-1474.117
Time:	22:50:17	BIC	-1452.041
Sample:	1	HQIC	-1465.530

	coef	std err	z	P>|z|	[0.025	0.975]
const	9.337e-05	0.001	0.125	0.900	-0.001	0.002
ar.L1.D.y	-0.1387	0.053	-2.598	0.009	-0.243	-0.034
ar.L2.D.y	-0.1660	0.047	-3.499	0.000	-0.259	-0.073
ma.L1.D.y	-0.6660	0.040	-16.842	0.000	-0.744	-0.588

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.4178	-2.4183j	2.4541	-0.2772
AR.2	-0.4178	+2.4183j	2.4541	0.2772
MA.1	1.5015	+0.0000j	1.5015	0.0000

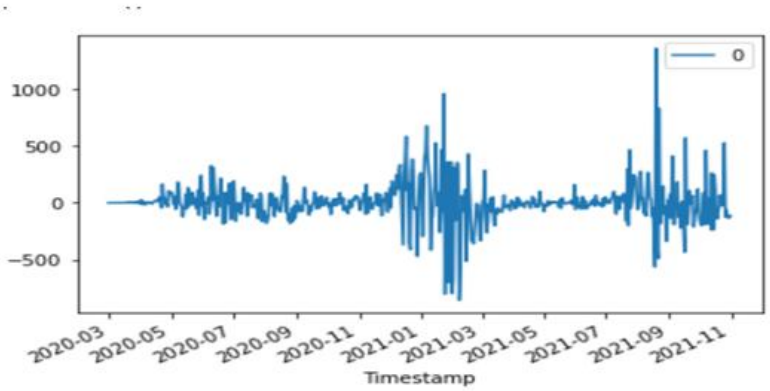
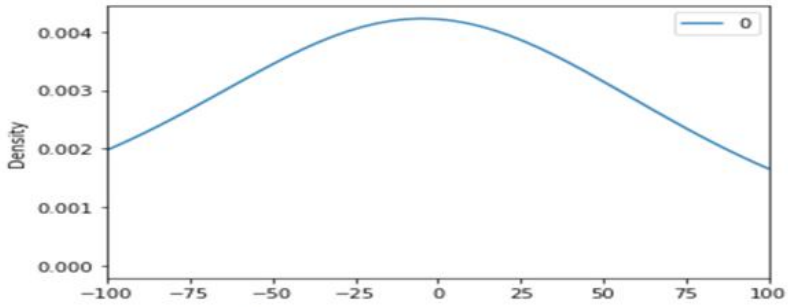


Figure 6: Residual Plot

Figure 6 shows the residuals of the data, and we can observe that most of the data is distributed around zero, which is expected as the error term being white noise and thus making the model estimate robust.

```
print(residuals.describe())
```




```
                0
count   611.000000
mean     -0.004976
std     176.997380
min    -858.558849
25%     -53.184301
50%      -3.676694
75%      46.670273
max    1359.624798
```

Figure 7: Residual Normality Plot

From Figure 7, the ARIMA model residual plot is observed to have a similar shape with the normal distribution. This keeps the normality of the residual assumptions for the model estimate to be reliable. The results are well fitted and most of the points are superimposed on each other.

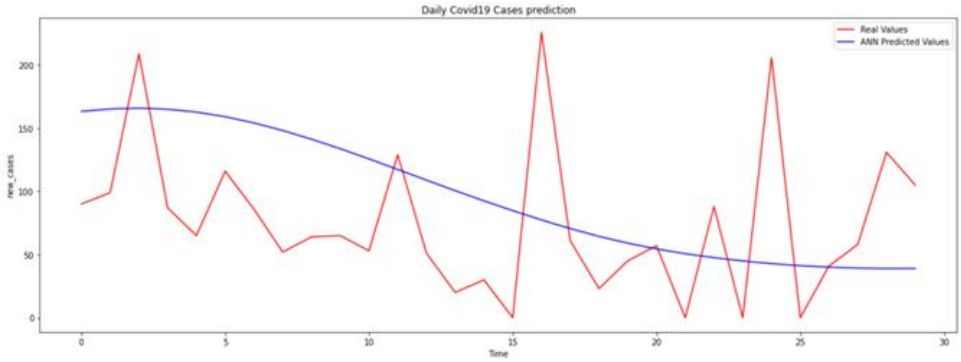


Figure 8: ANN Prediction Plot

From the Figure 8 above, it is observed that the predictions do not well suit the real values.

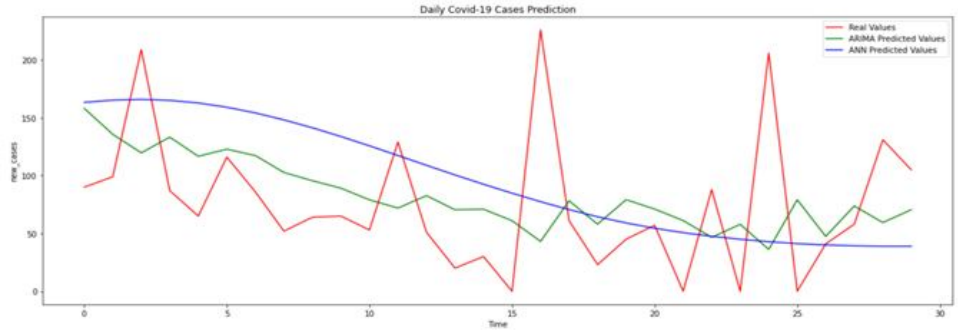


Figure 9: Models Prediction Plot for ARIMA, ANN and Real Values

From Figure 9, it can be inferred that the ARIMA model appears to fit better than ANN.

Table 7: Prediction Table for ARIMA and ANN Model

Date	Confirmed Cases	ARIMA Prediction	ANN Prediction
1/11/2021	90	158	163
2/11/2021	99	136	165
3/11/2021	209	120	166
4/11/2021	87	133	165
5/11/2021	65	117	163
6/11/2021	116	123	159
7/11/2021	86	117	154
8/11/2021	52	103	148
9/11/2021	64	96	141
10/11/2021	65	89	134
11/11/2021	53	79	126
12/11/2021	129	72	117
13/11/2021	51	83	109

14/11/2021	20	70	101
15/11/2021	30	71	93
16/11/2021	0	61	85
17/11/2021	226	43	77
18/11/2021	61	78	71
19/11/2021	23	58	64
20/11/2021	45	79	59
21/11/2021	57	71	54
22/11/2021	0	61	51
23/11/2021	88	47	48
24/11/2021	0	58	45
25/11/2021	206	36	43
26/11/2021	0	79	41
27/11/2021	41	47	40
28/11/2021	58	74	39
29/11/2021	131	59	39
30/11/2021	105	71	39

Table 8:ARIMA and ANNModel Evaluation Techniques

Finite Model Properties	MSE	RMSE	MAE
ARIMA	4123.910982	64.21768434	50.83978733
ANN	5066.314303	71.1780465	60.56677167

From Table 8,the RMSE value of ARIMA model, 64.217 is less than the RMSE value of ANN model, 71.178. This pattern of differences is consistent with the MSE and MAE, where the ARIMA model is observed to have returned the least values in all cases. Therefore, the smaller the finite model property, the better is themodel.

6.0Conclusion

In this study, comparison of prediction accuracy between the linear time series model (ARIMA) and the nonlinear time series model (ANNs) has been conducted. The prediction of the daily confirmed COVID-19 cases in Nigeria was carried out from its index case, February 28, 2020 till November 30, 2021.

The evaluation criteria for ARIMA and ANNs were compared and the results revealed that the conventional ARIMA model outperformed ANN in terms of minimum prediction error and forecasting ability.

The ANN model used forward propagation algorithm, which is by extension Recurrent Neural Networks (RNN) with three units in the hidden layer, two lags and the learning rate equal to 0.1. This is not the best fit for the Nigeria COVID-19 daily confirmed cases.

The ARIMA (2,1,1) is therefore the best fitted model for the daily confirmed COVID-19 cases in Nigeria among other Box Jenkins models since it outperformed the ANN model.

References

Adhikari, S.P., Meng, S., Wu, Y.J., Mao, Y.P., Ye, R.X., Wang, Q. Z., Sun, C., Sulvia, S., Rozelle, S. and Raat, H. (2020). Epidemiology, Causes, Clinical Manifestation and Diagnosis , Prevention and Control of Coronavirus Disease During the early Outbreak Period: A Scoping Review, Infect. Dis. Poverty, 9, 29.

Ajao, I.O. , Awogbemi, C.A. and Ilugbusi, A.O. (2020). Vector Autoregressive Models for Multivariate Time Series Analysis on COVID-19 Pandemics in Nigeria. Journal of Biology and Medical Research, 3(2), pp 171-181.

Akeyede, H.R., Bakari, H.R. and Muhammad, R.R. (2022). Robustness of ARIMA and ACP Models to Over Dispersion in Analysis of Count Data. Journal of the Nigerian Statistical Association, 34, pp 11- 21

Box, G.E., Jenkins, G. M., Reinsel, G. C. (1994). Time series analysis: Forecasting and Control, 3rd ed., New Jersey: Prentice Hall.

Ceylan, Z. (2020). Estimation of CoVID-19 Prevalence in Italy, Spain and France. Science of Environment, 729, 1388.

Dilbag, S., Kamar, V. and Manjit, K. (2020). Classification of COVID-19 Patients from Chest CT Images using Multi-objective Differential Evolution Based Convolutional Neural Networks. European Journal of Clinical Microbiology and Infectious Diseases, 39, pp 1379 -1389.

Fang, P.H. and Berlin, B.W. (2014). Using the Transfer Function Mode in Analyzing and Forecasting Students Study Achievement. Journal of Business and Economics, 5(1), pp 2052- 2056.

Fanelli, D. and Pizza, F. (2021). Analysis and Forecast of COVID -19. Spreading Chaos Solitons and Fractals, 134, pp1-5.

Gupta, R. and Pal, S.K. (2020). Trend Analysis and Forecasting of COVID-19 Outbreak in India. Med Rx, pp 1-19.

Hyup, R. (2007). Forecasting Volatility of Price Index. Expert Systems with Applications, 33(4), pp 916-922.

Khashei, S., Bijari, M. and Adali, G.A.R. (2009). Improvement of Autoregressive Integrated Moving Average Models using Fuzzy Logic and Artificial Neural Networks. Neuro Computing, 4(72), pp 956- 967.

Liao, Z. and Wang, J. (2010). Forecasting Model of Global Stock Index by Stochastic Time Effective Neural Network. Expert Systems with Applications, 37(1), pp 834-841.

Mamoud, K.O., Assem, A.Y. (2013). Comparison between ARIMA Models and Artificial Neural Networks in Forecasting Al-Quds Indices of Palestine Stock Exchange Market. The 25th International Conference on Statistics and Modeling in Human and Social Sciences, Department of Statistics, Faculty of Economics and Political Sciences, Cairo Uiversity, 25, pp 1-24.

Nawaf, N.H., Wagar, A.K., Wagar, A. Samer, H.A., Hyas, K. and Banvar, N.H. (2021). Artificial Neural Networks for Prediction of COVID-19 in Saudi Arabia. Computers, Materials and Continua, Tech Science Press, 16(3), pp 2787- 2796.

Nwafor, G.O., Iwu, H.C., Anyasodo, U.N. (2022). Transfer Function Modelling of COVID-19 Pandemic in Nigeria. Journal of the Nigerian Statistical Association, 34, pp 74-87.

Piccolomini, E.L and Zama, F. (2020). Monitoring Italian COVID-19 Spread by a forced SEIRD Model. Plos One, 15(8), pp 1-17.

Pullano, G., Pinotti, F, Valdona, E., Boelle, P.Y, Polletto, C. and Colizza, V. (2020). Novel Corona Virus (2019-nCov) Early State Importation Risk to Europe. Euro Surveillanceb, 25(4), pp 1560- 7917.

Rediat, T. (2020). Stochastic for Predicting COVD-19Prevalence in East African Countries. Infectious Disease Modeling, 5, pp 598 -607.

Samsuddeen, S. and Muhammad, S. (2020). Application of ARIMA and Artificial Neural Networks Models for Daily Cumulative Confirmed COVID-19 Prediction in Nigeria. Equity Journal of Science and Technology, 7(2), pp 83-90.

Tabachnick, B.G., Fidell, L.S. (2001). Using Multivariate Statistics Pearson Education, 4th ed.,Upper Saddle River , NJ, USA.

Unvan, Y.A. and Demirel. O. (2020). Diseases, Forecasting the number of Cases and Deaths in Turkey. Medical Science and Discovery, 7(61), pp 535- 543.

Wang, J.J., Wang, J.Z., Zhang, Z.G., Guo, S.P. (2012). Stock Index Forecasting Based on a Hybrid Model, Omega, 40(6), pp 758 – 766.

WHO (2020). Coronavirus disease 2019. *World Health* Organization; https://doi.org/10.1001/jama.2020.2633.