A Machine Learning Univariate Time series Model for Forecasting COVID-19 Confirmed Cases: A Pilot Study in Botswana

Ofaletse Mphale[1†], Ezekiel U Okike[2††]and Neo Rafifing[3†††] ,
Ofaletse_offie@hotmail.com euokike@gmail.com neorafifing@yahoo.com
University of Botswana, Department of Computer science, Gaborone, Botswana

Summary
The recent outbreak of corona virus (COVID-19) infectious disease had made its forecasting critical cornerstones in most scientific studies. This study adopts a machine learning based time series model - Auto Regressive Integrated Moving Average (ARIMA) model to forecast COVID-19 confirmed cases in Botswana over 60 days period. Findings of the study show that COVID-19 confirmed cases in Botswana are steadily rising in a steep upward trend with random fluctuations. This trend can also be described effectively using an additive model when scrutinized in Seasonal Trend Decomposition method by Loess. In selecting the best fit ARIMA model, a Grid Search Algorithm was developed with python language and was used to optimize an Akaike Information Criterion (AIC) metric. The best fit ARIMA model was determined at ARIMA (5, 1, 1), which depicted the least AIC score of 3885.091. Results of the study proved that ARIMA model can be useful in generating reliable and volatile forecasts that can used to guide on understanding of the future spread of infectious diseases or pandemics. Most significantly, findings of the study are expected to raise social awareness to disease monitoring institutions and government regulatory bodies where it can be used to support strategic health decisions and initiate policy improvement for better management of the COVID-19 pandemic.

Key words:
COVID-19, Coronavirus, ARIMA, Box-Jenkin, Time series, Machine learning, ACF, PACF, AIC

1. Introduction

Coronaviruses are infectious diseases that are closely related to common cold, Middle East Respiratory Syndrome coronavirus (MERs) and Severe Acute Respiratory Syndrome coronavirus (SARs). These diseases are once known to diffuse from animals to human beings. For example; SARs, was known to transfuse from civet cats to humans while MERs was transmitted to humans from a type of camel [1]. The corona virus was officially termed as "COVID-19" by the World Health Organization (WHO) and its first incident was registered in Wuhan city in China on December 2019. Since then, the virus had spread rapidly reaching different region of the world [2-4].

Fatalities from COVID-19 had been presented in amplifying figures globally. Empirical studies show that fatalities due to COVID-19 had surpassed 1.9 million, with confirmed cases exceeding of 88.5 million worldwide [5] [6]. In state-of-art, studies had shown that there had been limited attempts in clinical trials conducted to evaluate potential COVID-19 treatments [7],[8]. However, some crucial recovery measures had been outlined by WHO [6]such as; self-isolation, drinking of plenty of water, consumption of paracentamol and adequate rests

COVID-19 disease can affect individuals in different age spectrums. It is transfused through human to human contact. Its symptoms are characterized by diseases like flu, fever, fatigue and respiratory complication. Elderly people with other chronic diseases like diabetes and high blood pressure are the most vulnerable to the undesirable effects of the disease. Some safe guard measures such as frequent hand washing, wearing face mask and social distancing had been suggested in some studies to reduce contamination with the disease[6] , [9]. The consequences of infectious diseases do not only detriment human health, but it is also an economic burden. While most countries are employing procedures to control the virus such as lock down, mobility restrictions, quarantine and more, the accurate prediction of infectious diseases remains a difficult task [10] because, the forms and outbreaks of infectious diseases are often unknown [6].

Time series models are popular machine learning techniques applied in different scientific grounds to discover trends and relationships in series data. This study adopts ARIMA model to forecast corona virus confirmed cases in Botswana over 60 days period. ARIMA model follows a Box-Jenkin approach for time series forecasting. Results of the study are expected to raise social awareness to disease monitoring institutions and the government regulation bodies where it could be utilized to support strategic health decisions and enhance policy improvement procedures, for better management of the COVID-19 disease.

The entire paper is organized as follows; Section 2 presents the Literature review of the subject being studied. That is different theoretical and empirical scholars' perceptions on application of various time series models in forecasting of infectious diseases. In Section 3, Methodology framework to be followed by the study analysis process is presented. Section 4 presents Results

Manuscript received December 5, 2021
Manuscript revised December 20, 2021
https://doi.org/10.22937/IJCSNS.2022.22.1.31

and Discussion of the study. Finally Conclusions and Future works are discussed in section 5.

2. Literature Review

With the recent improvements of forecasting methods like machine learning, artificial intelligence and mathematical models, scholars had come to appreciate them and integrate them in different studies to tackle real world tasks. Predictive analytics learn from historical data and utilizes machine learning approaches to stem future conclusions. The application of machine learning algorithms in technical grounds like engineering, computer science, medicine, statistics etc. had made it possible to recognize infectious diseases patterns, accelerate diagnosis and to efficiently forecast their future directions.

Infectious diseases are caused by pathogenic microorganisms such as bacteria, viruses, parasites or fungi which are diffused between individuals or an animal [11]. Zoonotic diseases are groups of infectious diseases that affect animals, but can cause diseases when transmitted to humans [12]. Studies had shown that to-date, various models and tools had been developed to predict and monitor outbreaks of infectious diseases. For instance; Chaurasia and Pal [13]compared different forecasting methods such as Holt linear trend method, naive method, single exponential smoothing, simple average, Holt-Winters method, moving average and ARIMA using root mean square error score. In their findings it was concluded that the naïve model outperformed all other models. However, the ARIMA model was nominated by grid search method as the best fit model fit for the data. Findings further established that the number of COVID-19 deaths will surpass 600,000 in January 2021.

Abuhasel, Khadr and Alquraish [14] applied classical SIR model to predict the highest number of COVID-19 cases that could be recognized in order to flatten the curve. Furthermore, ARIMA model was used to predict the prevalence cases of COVID-19 infections. In their findings the SIR model affirmed that the containment technique used by Saudi Arabia to minimize the spread of the disease was effective. In evaluating the performance of the models, ARIMA proved to be a good forecasting method from current data.

In another study, Chae, Kwon and Lee [15]applied and optimized various deep learning algorithms to predict outbreak of infectious diseases such as chicken pox using social media big data. Predictive models such as ARIMA, deep neural network (DNN) and long-short term-memory (LSTM) were critically compared. Based on the results, it was established that DNN and LSTM models outperformed the ARIMA model. However, LSTM model produced more reliable predictions when compared to DNN model especially when modelling the disease outbreak.

Jia *et al*. [10] applied and compared mathematical models such as Logistic model, Bertalanffy model and Gompertz model on the SARs pandemic dataset. In their findings it was established that the three models performed differently with different parameters in different regions. Formerly, Wang *et al*. [16]evaluated performance of forecast models such as ARIMA, LSTM, back-propagation artificial neural network (BP-ANN) and Seasonal Trend Decomposition method by Loess + ARIMA using malaria data from Yunnan Province in China. In their findings it was shown that the four models performed better when stacked with gradient-boosting regression trees. In their study it was also established that assemble algorithms could improve prediction of the infectious diseases prediction models.

In another study Wang *et al*. [17] , scrutinized the seasonal patterns of hand, foot and mouth disease (HFMD) in children in Mainland, China. The study compared prediction performance of the LSTM, ARIMA and auto-regressive neural network models. In their findings it was concluded that LSTM model was a best fit for the data and its prediction performance outperformed other models. The results further showed that trend of HFMD cases is rising in summer signifying high-risk season. Moreover, results showed that the LSTM method was relevant in predicting outbreak of malaria disease cases.

3. Methodology

3.1 Dataset Description

The dataset used in the study was acquired from [18] and other relevant sources [5], [19], [20]. These are data repositories which are freely available for public use for academic and non-academic purposes. The acquired dataset consisted of COVID-19 global registered cases (Confirmed, Recovered, Deaths) from 31st November 2019 to 12th January 2021. Since the study was only interested in investigating the spread of COVID-19 confirmed cases in Botswana, some observations and attributes were pruned from the dataset. The IBM Statistical Package for the Social Sciences (SPSS) and Python Jupyter notebook were used to prepare, transform and analyze the dataset. Therefore the final dataset comprised of only observations of confirmed cases in Botswana registered from month of 4th April 2020 to month of 12th January 2021. The proposed methodology framework followed by the study analysis process is illustrated in Figure 1 as shown.

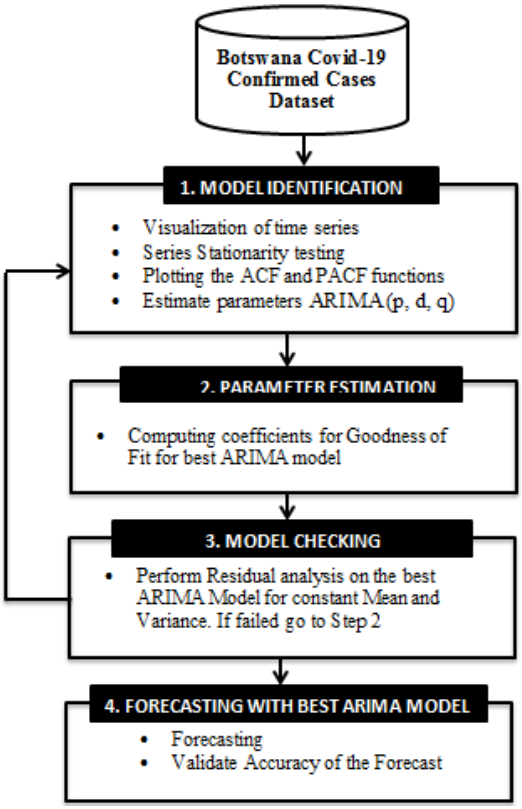

Fig. 1 Proposed methodology framework followed by the study analysis process

3.2 Time Series Analysis

Time series is simply expressed as a set of data points ordered in time [21]. It assumed to comprise of a random walk (non-stationary series) and white noise (zero mean stationary series). There are two main classes of time series forecasting approaches which are; univariate time series forecasting and multivariate series forecasting. In univariate time series forecasting, predictions of future data points primarily depends on previous values in the series to generate forecasts whereas in multivariate time series analysis other predictors other than the series values(exogenous variables) are taken in to account in generating forecasts. In this study, based on the structure and nature of the dataset being examined, a univariate time series analysis approach was deemed appropriate to model forecasts.

In the step,the study needs to establish whether the series can be decomposed using additive method or multiplicative method. In an additive time series analysis, seasonality and residuals are independent of the trend whereas in multiplicative time series is vice versa. Mathematically an additive time series method is given in equation (1).

$$O_t = T_t + S_t + R_t \quad (1)$$

Where, O_t– represents the output, T_t– represents the trend, S_t– represents the seasonality and R_t – represents the output.

Alternatively, multiplicative time series is formerly described as shown:

$$O_t = T_t * S_t * R_t \quad (2)$$

In this paper, an additive time series decomposition method was chosen over multiplicative time series decomposition. Then, ARIMA model was developed and used to generate forecast of the daily confirmed cases in Botswana.

3.2 ARIMA Model

ARIMA model is a statistical machine learning based algorithm used in time series forecasting. It was discovered by George Box and Gwiltyn Jenkins and is known as Box-Jenkins model [22]. ARIMA model extends Auto Regressive (AR) and Moving Average (MA) by integrating with the order of differencing steps (I). It depends on the known historical data in order to establish forecast values [23].

To successfully implement the ARIMA model, a non-stationary time series must be transformed from random walk to white noise [24]. Random walk series produces unreliable and unstable forecasts. A Stationary series has a constant mean, variance and its autocorrelation structures are not affected by fluctuations of time. To test for stationarity of the series, an Augmented Duckey Fuller (ADF) test can be used [25].An ADF test examines the null hypothesis for the presence of a unit root in series which is recognised at value of alpha (α) = 1. The criterion used to test the null hypothesis for presence of the unit root is given as shown:

Given α is = 1,
Then,

H₀: The series has a unit root. The series is non stationary.

H₁: The series does not have a unit root. The series is stationary.

To determine the presence of unit a root in series, the *p-value* obtained must be less than the significance level threshold of 5%. This implies that the null hypothesis is

rejected and the alternative hypothesis is accepted, indicating that the series is indeed stationary. However, if the *p-value* score is greater than the significant level threshold of 5%, the null hypothesis is accepted indicating that the series is non stationary.

The formal definition of ARIMA model is given as shown in equation (3):

$$ARIMA\ (p, d, q) \tag{3}$$

Where ***p*** represents the order of AR polynomial indicating of the autoregressive model lags, *d* represents the order of the differencing steps, *q* represents the MA polynomial order of the moving-average process.

If the series is already stationary, then ARIMA model can be presented as an *ARMA (p, q)* with a differencing sequence of *d* times, where $p, d, q \geq 0$. It can be simplified as ARMA model as shown:

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-1} + a_t - \sum_{j=1}^{q} \theta_i a_{t-j,} \tag{4}$$

Where $\phi_{1,} \ldots, \phi_p$ are the AR parameters to be estimated, $\theta_{1,} \ldots, \theta_q$ are the MA parameters to be estimated and $a_{1,} \ldots, a_t$ are a series residuals that follows a normal distribution. The equation (4) can also be simplified by applying the Box-Jenkins backshift operator as shown:

$$\left(1 - \sum_{i=1}^{p} \phi_i \beta^i\right) Y_t = \left(1 - \sum_{j=1}^{q} \theta_j \beta^j\right) a_{t,} \tag{5}$$

Alternatively, equation (5) is further condensed to equation (6) as shown:

$$\phi_p(\beta) Y_t = \theta_q(\beta) a_{t\,,} \tag{6}$$

Where,

$$\phi_p(\beta) = \left(1 - \sum_{i=1}^{p} \phi_i \beta^i\right) \text{and } \theta_q(\beta) = \left(1 - \sum_{j=1}^{q} \theta_j \beta^j\right)$$

If the series is non-stationary, then ARIMA model can be extended by integrating with the order of differencing steps as illustrated in equation (7):

$$\begin{aligned} W_t &= Y_t - Y_{t-1} = (1 - \beta)\, Y_t \\ W_t - W_{t-1} &= Y_t - 2Y_{t-1} + Y_{t-2} \\ &= (1 - \beta)^2 Y_t \\ &\vdots \\ W_t - \sum_{k=1}^{d} W_{t-k} &= (1 - \beta)^d Y_{t,} \end{aligned} \tag{7}$$

Where, *d* is the order of differencing steps. Then, replacing the Y_t in the ARMA model with the differences defined in equation (7), ARIMA *(p, d, q)* model is simplified as shown in equation (8).

$$\phi_p(1 - \beta)^d Y_{t,} = \theta_q(\beta) a_{t\,,} \tag{8}$$

3.3 Model testing for goodness of fit

Model testing for goodness of fit validates whether ARIMA model is an appropriate fit to the data. In this study, a Grid Search Algorithm (GSA) was developed with python language and was used to optimize AIC metric for evaluating different permutations of ARIMA models. The lowest AIC score was used to denote a good model fit for the data. Mathematically AIC is given as shown:

$$AIC = -2\ (log\text{-}likelihood) + 2k \tag{15}$$

Where, *k* is the number of model parameters (the number of variables in the model plus the intercept), *log-likelihood* is a measure of model fit. The higher the *log-likelihood* measure, the better the model fit the data.

3.4 Evaluation of accuracy of estimated forecasts

There are various error metrics that can be applied to evaluate the accuracy of the forecasts errors in ARIMA models. In this study Ljung-Box accuracy metric was applied in successions on residuals at different time lag intervals. Then, different measures of the residuals errors were examined and used to determine various conclusions.

4. Results and Discussions

The major objective of the study was to model forecasts of COVID-19 confirmed cases in Botswana using ARIMA model. Therefore in order to develop the most suitable and reliable ARIMA model, the study followed the four main stages which includes; model identification, model parameter estimation, diagnostic checking and forecasting with best ARIMA model. These stages are also given in Figure 1 – methodology framework.

4.1 Model Identification

In this phase, the series for COVID-19 confirmed cases was plotted and its various components such as seasonality, trends and noise were examined. Plotting the series is critical for better understanding of the series structure in order to determine the appropriate forecasting model for the data [14]. Figure 2 presents the graphical

illustration of the COVID-19 confirmed cases in Botswana from month of 4th April 2020 to 12th January 2021.

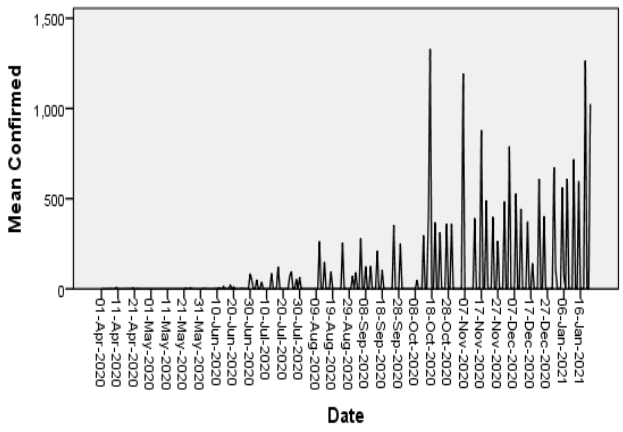

Fig. 2 COVID-19 confirmed cases in Botswana from month of 4th April 2020 to 12th January 2021.

Fig. 2 shows that from April 2020 to January 2021, COVID-19 confirmed cases in Botswana were rising steadily with a horizontal trend and some periodic spikes. The series also depicts components of weak stationarity series which can be described using an additive model. From period of the month of April 2020 up to early October 2020 results showed that COVID-19 confirmed cases were at minimum. These results can be related to the effectiveness of government policies and precaution measures which were imposed during the time e.g. national lock down, quarantine, restriction of movements, compulsory wearing of face masks in public areas and others.

In late October 2020 the series depicts its sharpest spike of numbers of COVID-19 confirmed cases. These results can be related to the early period of state of emergency extension, when most precaution measures were still under consideration or were not fully implemented. However the confirmed cases degraded towards the month of December 2020. Then, rise towards the month of January 2021. This could also suggest that the government precaution measures which were imposed during that period such as curfew, regulation of liquor stores trading hours, social distancing, prohibition of public gatherings and more, had a slight impact towards controlling the spread of COVID-19 daily infections.

To validate whether the ARIMA model was a relevant model for the data, the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots were fitted with the data. The behavioral patterns of the correlograms were examined in respective ACF and PACF plots. Figure 3 presents the results of fitting confirmed dataset in ACF and PACF plots.

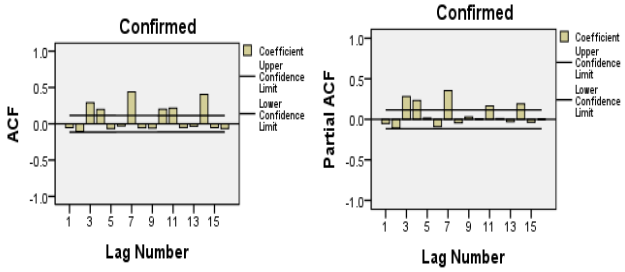

Fig. 3 Results of ACF and PACF correlograms plots fitted in confirmed cases dataset

Fig 3 shows that the both ACF and PACF correlograms follow the same pattern of significance at lag 3, 4, 7 and 11 respectively. This indicates a series can be modelled appropriately using an ARIMA model processes. To gain better understanding on the series trend and its seasonal properties, we decomposed the series following an additive model in Seasonal Trend Decomposition method by Loess (STL). Figure 4 depicts the results of series decomposition using the STL method. Results show that the trend COVID-19 confirmed cases in Botswana is constantly rising in a fluctuating upward trend over time. The series also show daily cyclic patterns of COVID-19 confirmed cases over time, with no clear seasonal trend.

In the residual plot, the series show that residues had been constant in the early months of the series. That is from the beginning of April 2020 until September 2020. However, after October 2020 there are variations in residuals depicting a random noise and characteristics of non stationarity series. In a stationary time series, the residual distribution are assumed to compose of variance that revert around zero mean [25].

Fig. 4 Results of STL series decomposition of COVID-19 Confirmed cases in Botswana over Time

4.2 Testing for Stationary

The study needed to confirm whether the series was stationary or not. In that way an ADF test method was applied to examine the null hypothesis for the presence of unit root and the *p-value* measure obtained was 0.330804. This was greater than the threshold significant level of 5%, and the study failed to reject the null hypothesis, affirming the series was not stationary. To effectively model and generate forecasts with ARIMA models, a non-stationary series need to be converted to white noise. Therefore different combination of ARIMA models parameters were fitted to the data using GSA to achieve a white noise (See Section 4.3)

4.3 Model Parameter Estimation

In this phase, the variance in the series was normalized by initially applying a log transformation process. Then GSA was run recursively on different permutations of ARIMA model parameters in range between 0 and 6. This was to ensure the model would not over fit the dataset. The ARIMA models selection for goodness of fit was based on AIC metric. In this study the lowest score of AIC metric was used to determine the best fit ARIMA model at ARIMA (5, 1, 1) and the AIC measure of 3885.091 (See Fig. 5).Having established the best fit ARIMA model, the study also needed to confirm that the chosen model was appropriate for generating reliable forecasts and residuals had no serial correlation with the data. Therefore model checking phase was conducted.

ARIMA Models	AIC Scores
ARIMA(2,1,2) with drift	3887.532
ARIMA(0,1,0) with drift	4151.456
ARIMA(1,1,0) with drift	4083.781
ARIMA(0,1,1) with drift	3916.383
ARIMA(0,1,0)	4149.502
ARIMA(1,1,2) with drift	3902.917
ARIMA(2,1,1) with drift	3887.491
ARIMA(1,1,1) with drift	3909.989
ARIMA(2,1,0) with drift	3971.975
ARIMA(3,1,1) with drift	3886.553
ARIMA(3,1,0) with drift	3925.584
ARIMA(4,1,1) with drift	3884.729
ARIMA(4,1,0) with drift	3925.421
ARIMA(5,1,1) with drift	3885.091
ARIMA(4,1,2) with drift	3886.184
ARIMA(3,1,2) with drift	3887.049
ARIMA(5,1,0) with drift	3928.319
ARIMA(5,1,2) with drift	3885.156
ARIMA(4,1,1)	3888.293

Fig. 5 Results of fitting different ARIMA models and corresponding AIC scores using GSA

4.4 Model Checking

An ADF test was applied on the chosen ARIMA model to test examine the behavior of residuals around it, and the *p – value* score obtained was 4.193195e-14. This was less than the threshold critical value of 5%., which implied that the series was indeed stationary. The study also needed to examine the stability of chosen model in generating forecasts, therefore the behavior of residuals around the model were analyzed using the ACF and PACF plots. Figure 6 presents the results of residual analysis in ACF and PACF plots.

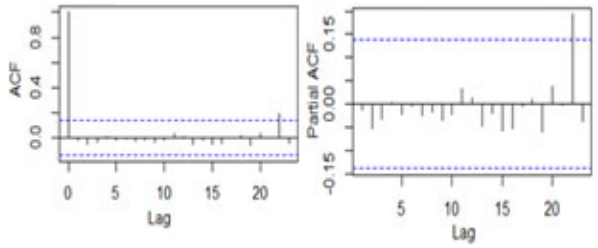

Fig. 6 Results of ACF and PACF residual analysis plots

Fig.6 shows that residuals were normally distributed with constant variance and mean over time. However there were some significant spikes observed at lag 0 and lag 22 respectively indicating a slight auto correlation of residues with the model. To validate whether the significant spikes would have major impact towards the forecast intervals, a Ljung-Box test was executed on the chosen model and the results showed that the *p-value* measure was 0.7814. This was more than the critical value of 5%, which indicated that residuals were purely random and depicted no significant autocorrelation with the model. Moreover, the results conveyed that residuals were a white noise, which affirmed that the chosen model was an adequate fit to the data. Therefore, with the derived assumptions the model was considered for generating reliable forecasts of COVID-19 confirmed cases in Botswana over 60 days period.

4.5 Forecasting with best ARIMA Model

The chosen ARIMA model (5, 1, 1) was used to generate forecasts of COVID-19 confirmed cases forecast over 60 days period. That is from the month of 22ndJanuary 2021 to March 22nd 2021. The actual values of confirmed cases (observed) were plotted against the forecasted values. The 95% standard error bond lines (UCL and LCL) were also plotted to guide the forecasts error limits. Figure 7 depicts the results of plotting the observed and forecast values of COVID-19 confirmed cases in Botswana over 60 days period. In Fig. 7, observed

values are represented by a red line, the dotted line represents the error bond lines and the dark blue line depicts the trend of the forecasts values.

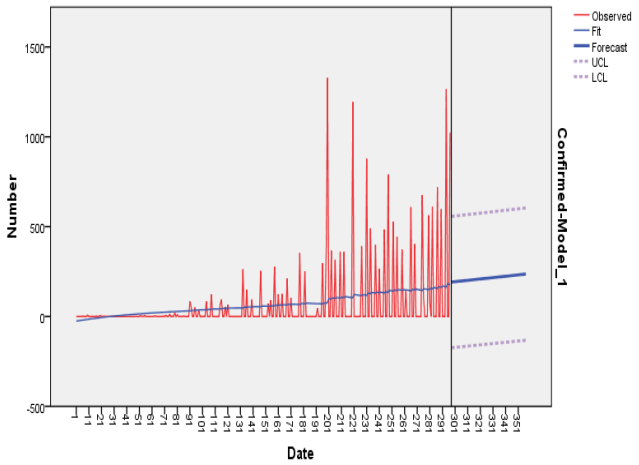

Fig. 7 Observed values vs. Forecast values of COVID-19 confirmed cases in Botswana over the next 60 days period

Fig. 7 shows that confirmed cases in Botswana are expected to rise steadily in an upward trend with random fluctuation over the next 60 days period. These results are valid provided the environmental variables are constant, i.e. the government precaution measures, policies and other strategies used to control the virus during the time. To validate the accuracy of the forecasts, a series of Ljung-Box tests were performed on the forecasts residual errors at different lag intervals. The *p-value* measure was used to evaluate residuals independence and randomness around the generated forecasts. Table 1 depicts the results of conducting successions of Ljung-Box tests on forecasts residuals.

Table 1: Ljung-Box test results on confirmed cases forecasts residuals

Number of Lag runs	***Ljung-Box test (p- value results)***
5	0.9728
10	0.9995
15	0.9997
20	0.9799
25	0.9689

Table 1 shows that residuals of the forecasts errors are normally distributed with mean close to zero and the *p-value* measures is greater significant confidence interval of 5%. In a good forecast model, the assumption is that residuals should resemble zero correlations over time. Furthermore, Ljung Box test assumes that residuals are not significant if the *p-value* is greater than zero mean [23]. Therefore, based on the derived assumptions, it was concluded that the generated forecasts were reliable and volatile in providing guidance on understanding the direction of the future spread of COVID-19 confirmed cases in Botswana. In addition, results suggested that much effort will be required from the government of Botswana in order to flatten the curve of the pandemic. Therefore, it was recommended that stricter precaution policies and measures should be imposed by the government for appropriate management of the pandemic. For example; an execution of a second national lock down, stringent curfew regulations, stringent penalties towards of compulsory wear of face mask in public areas, strict social distancing regulations and the public should also abide by COVID-19 precaution measures at all the times.

5. Conclusion and Future works

The accurate forecasting of COVID-19 infectious disease had become critical for the stability of every country's economy and societal wellbeing. This study adopted an ARIMA model to forecast confirmed cases in Botswana for the next 60 days period. Findings of the pilot study suggest that ARIMA model is a powerful tool that can be used to generate reliable forecasts which can guide on estimating future discourse of pandemics. Whereas there are some forecasting limitations associated with ARIMA models, the reliability and volatility of forecasts errors can be improved. In some cases accuracy performance metrics like a Ljung Box test can be used.

In forecasting COVID-19 confirmed cases in Botswana over 60 days period, pilot findings suggest that provided the environmental variables remain constant (i.e. the current government precaution procedures, policies and other strategies to control the virus are imposed) then confirmed cases in any environment are expected to rise gradually following a steep upward trend with random fluctuations. Therefore in order to effectively manage the COVID-19 pandemic, the study recommends and support government imposition of stricter precaution policies and measures as and when necessary. This could involve execution of national lock downs, strict curfew regulations, more strict penalties towards of compulsory wearing of face mask in public areas, strict social distancing regulations and public acceptance and abiding by the COVID-19 precaution measures at all the times.

This study being one of the ground-breaking studies that models time series forecasts of COVID-19 infections in Botswana, findings are expected to support strategic management decisions and policy improvements towards appropriate management the COVID-19 disease in the country. In future research, the study intends to use a larger data set that covers longer periods in order to improve the reliability of the forecasts. Moreover,

accuracy metrics such as Root Mean Squared Error, Mean Absolute Percentage Error and Mean Absolute Error could be applied on the forecast errors to investigate the correlation behaviour of residuals. The study also intends adopt a Seasonal ARIMA model which takes in to account of some independent variables and compared to machine learning models like, artificial neural networks and Facebook prophet model to investigate different ways to correlate forecasts of COVID-19 confirmed cases.

References

[1] R. Singla, A. Mishra, and R. et al. Joshi, "Human animal interface of SARS-CoV-2 (COVID-19) transmission: a critical appraisal of scientific evidence," *Springer Nature Switzerland*, pp. 119–130 (2020).DOI:10.1007/s11259-020-09781-0, 2020.

[2] BBC News. (2020, December) Covid-19 pandemic: Tracking the global coronavirus outbreak. [Online]. https://www.bbc.com/news/world-51235105

[3] Our World in Data. (2020, January) Coronavirus Pandemic (COVID-19).[Online]. https://ourworldindata.org/coronavirus#citation

[4] V. Vara. (2020, April) Latest Analysis. [Online]. https://www.pharmaceutical-technology.com/features/coronavirus-outbreak-the-countries-affected/

[5] Worldometer. (2021, Aug.) Worldometer. [Online]. https://www.worldometers.info/coronavirus/

[6] WHO. (2021, January) WHO Coronavirus Disease (COVID-19) Dashboard. [Online]. https://covid19.who.int/

[7] IWK Health Center. (2021, July) Library Services. [Online]. https://library.nshealth.ca/COVID19Research/Publications

[8] S. H. Hodgson, K. Mansatta, and Mallet, G. et al, "What defines an efficacious COVID-19 vaccine? A review of the challenges assessing the clinical efficacy of vaccines against SARS-CoV-2," *The Lancert,*. DOI:10.1016/S1473-3099(20)30773-8, 2020.

[9] Z. Ceylan, "Estimation of COVID-19 prevalence in Italy, Spain, and France," *PMC US National Library of Medicine National Institute of Health*, DOI: 10.1016/j.scitotenv.2020.138817, 2020.

[10] L. Jia, K. Li, Y. Jiang, X. Guo, and T. zhao, "Prediction and analysis of Coronavirus Disease 2019," *NASA Astrophysics Data System*, [Online].https://arxiv.org/ftp/arxiv/papers/2003/2003.05447.pdf, 2020.

[11] A. Steptoe and L. Poole, *Infectious Diseases: Psychosocial Aspects*.: International Encyclopedia of the Social & Behavioral Sciences (Second Edition), 2015.

[12] World Health Organization. (2021, January) WHO Health Topic Page: Zoonoses. [Online]. https://www.who.int/topics/zoonoses/en/

[13] V. Chaurasia and S. Pal, "Application of machine learning time series analysis for prediction COVID-19 pandemic," *Springer*. DOI:10.1007/s42600-020-00105-4, 2020.

[14] K. A. Abuhasel, M. Khadr, and M. M. Alquraish, "Analyzing and forecasting COVID-19 pandemic in the Kingdom of Saudi Arabia using ARIMA SIR models," *Wiley - Computational Intelligence*, DOI: 10.1111/coin.12407, 2020.

[15] S. Chae, S. Kwon, and D. Lee, "Predicting Infectious Disease Using Deep Learning and Big Data," *International Journal of Environmental Research and Public Health.* DOI:10.3390/ijerph15081596, 2018.

[16] M Wang, H. Wang, J. Wang, and Lui et. al, "A novel model for malaria prediction based on ensemble algorithms," *PLoS ONE*, 14(12). DOI:10.1371/journal.pone.0226910, 2019.

[17] Y. Wang, C. Xu, S. Zhang, and Yang et. al, "Development and evaluation of a deep learning approach formodeling seasonality and trends in hand-foot-mouth disease incidence in mainland China," *Springer Nature*, 2019.

[18] Johns Hopkins University of Medicine. (2021, January) COVID-19 Dashboards by the Centerfor System Science and Engineering (CSSE) at Johns Hopkisn University. [Online]. https://coronavirus.jhu.edu/map.html

[19] Github. (2021, January) Github. [Online]. https://github.com/CSSEGISandData/COVID-19

[20] AccuWeather. (2021, January) Botswana Weather. [Online]. https://www.accuweather.com/en/bw/national/covid-19

[21] B. Fanoodi, B. Malmir, and F. F. Jahantigh, "Reducing demand uncertainty in platelet supply chain through artificial neural networks and ARIMA models," *Elsevier Computers in Biology and Medicine.*DOI: 10.1016/j.compbiomed.2019.103415, 2019.

[22] G.E.P. Box and G.M. Jenkins, *Time Series Analysis: Forecasting and Control, Revised Edition*. San Francisco: Holden Day, 1976.

[23] J. Fattah, L. Ezzine, and Z. Aman, "Forecasting of demand using ARIMA model ," *Internationa Journal of Business Management*, DOI:10.1177/1847979018808673, 2018.

[24] Z. Malkia, E Atlamb, E. Ewisc, and G etal Dagnewe, "ARIMA Models for Predicting the End of COVID-19 Pandemic and the Risk of a Second Rebound," *Research square*, DOI:10.21203/rs.3.rs-34702/v1, 2020.

[25] L. H. Koopmans, *Multivariate Spectral Models and Their Applications*. Science Direct - Elsevier, 1995.

Ofaletse Mphale received the Bachelor of Software Engineering from Multimedia University and MSc. in Computer Information Systems, from University of Botswana in 2012 and 2017, respectively. He has worked as a research assistant, teaching assistant and Computer demonstrator (from 2014), in the Department of Computer science in University of Botswana. His research interests include: Machine learning, Artificial intelligence, Time series analysis, Natural language processing, Big-data, Social media, Educational technology, ICT4D, Human Computer Interaction and Software project management. He is a member of Botswana Institute of Engineers, Botswana Qualification Authority Associates.

Ezekiel U. Okike received the B.Sc (Hons), MInf.Sc and PhD degrees, from the University of Ibadan, Nigeria in 1991, 1995 and 2007 respectively. He is currently a Senior Lecturer and Cluster Chair of Information Systems at the Department of Computer Science, University of Botswana.
He is a Senior Member of IEEE, and a Member of ACM. His teaching and research areas are in Software Engineering/Information Systems engineering, Information Systems Analysis and Design, Computer Organization and Architecture, Software Measurement and Models, Information/Cyber Security, Machine Learning and Learning Analytics.

Neo Rafifing received the BSc in Business Computing with Specialism in Management from Staffordshire University UK and MSc in Information Systems from Botswana International University of Science and Technology (BIUST). She has 10 years of experience in the IT industry.
IT Entrepreneurship Innovation, Block Chain Technology, E-Learning. E-government, Mobile Health, Internet of Things (IoT) and Adoption of IT technologies and Use. She has worked as a graduate teaching/research assistant at BIUST since 2014 in the computer science and information system department as well as business department. Her research interests include; Data Science and Business Analytics, Human Computer Interaction and Usability,