ARIMA AND ARIMAX IN THE MODELLING OF COVID-19 MORTALITY IN NIGERIA

Ajibode, I.A.* & Adeboye, N.O.
Department of Mathematics/Statistics,
Federal Polytechnic, Ilaro, Ogun state, Nigeria
*ilesanmi.ajibode@federalpolyilaro.edu.ng; nureni.adeboye@federalpolyilaro.edu.ng
***Corresponding author**

Abstract

The outbreak of COVID-19 disease has brought a lot of panic to the world. The number of deaths associated with COVID-19 greatly exceeds the other two corona viruses tagged severe acute respiratory syndrome corona virus-SARS-CoV and Middle East respiratory syndrome corona virus- MERS-CoV and the outbreak is still ongoing, which posed a huge threat to the global public health and mortality rate. Recent works have explored different modelling techniques including the popular box jenking technique, however, none of the modelling techniques was able to compare the predictive ability of the models. Thus, this paper compared the predictive ability of ARIMA and ARIMAX at predicting death due to COVID-19 in Nigeria. Data were extracted from the repository website of NCDC for 43 weeks on the number of confirmed cases and number of deaths. The original data was made stationary by differencing and appropriate models were developed. ARIMA (1,1,1,) and ARIMAX (1,1,1) were found to be the best model orders that correctly predict death due to COVID-19. However, it was discovered that ARIMAX (1,1,1) outperformed ARIMA (1,1,1,) based on its lowest MAE and RMSE of 5.9774 and 8.538211 respectively.

Keywords: COVID-19, ARIMA, ARIMAX, Stationary, Comparative

1.0 Introduction

The outbreak of Coronavirus in Wuhan, China, has brought serious panic and agony to nations around the globe. This virus is popularly referred to as 2019-nCoV, with an indication that it belongs the Orthocoronavirinae subfamily. It was first discovered in December. A large number of people have experienced and died of COVID-19 in China which was acclaimed to be the source of the virus and other countries of the world. However, the spread of the virus is still a mystery,

though extant literatures have claimed its transmission from animal to the human and from human to human (Lu et al, (2020), Ji et al (2020)). World Health Organization (WHO) declared that COVID-19 spread is just like that of H1N1 (2009), Polio (2014), Ebola in West Africa (2014), Zika (2016), and Ebola in the Democratic Republic of Congo (2019), Yoo (2019). WHO, has claimed that the disease can be spread through nearby interaction, by tiny beads formed through coughing, sneezing, or speaking (Control and Prevention, 2020; European Centre for Disease Prevention and Control, 2020; World Health Organization, 2020; Centers for Disease). Also, individuals could often turn out to be contaminated through being in contact with an affected exterior (World Health Organization, 2020; Centers for Disease Control and Prevention, 2020).

Friday January 28, 2020 marks the emergence of the virus in Nigeria since then of the virus was announced in Nigeria and the number of cases and deaths is on the rise every day. This has led to the inauguration of the country's National Coronavirus Emergency Operation Centre (Adepoju, 2020). This presence of the center can be felt in all states of the federation with the total number number of confirmed cases as at November 18, 2020 been 65,693, discharged cases are 61,457 and deaths is 1,163 (NCDC Nigeria, 2020; Wikipedia, 2020).

The pandemic is a serious challenge to the word even the world powers because it has affected both economic and social lives of people with emergence of new normal in daily activities.

Majority of research publications focused more on the epidemiology, trend analysis and forecasting for different cities and countries. These studies have presented long-term and short-term trend using time series data using various time series techniques. In the work of Li et al. (2020), he built a method for forecasting the ongoing trend with data-driven analysis and estimating the COVID-19 outburst size in China. Fanelli and Piazza (2020) studied the COVID-

19 pandemic temporal dynamics in mainland China, Italy and France. Roda et al. (2020) correlated the standard SIR and SEIR frameworks to model COVID-19 in Wuhan China. Wei et al. (2016) forecast the national and global spread of COVID-19 to determine the impact of the metropolitan-wide isolation of Wuhan and its neighbors. Wang et al. (2020) established the Patient Information Based Algorithm for evaluating the demise rate of COVID-19 in real-time by utilizing openly accessible datasets.

In the recent past, statistical and time series models have also been used to model and predict the prevalence of this pandemic. This is evidenced in the work of Ayinde et al. (2020), in which they subjected the COVID-19 cumulative confirmed cases in Nigeria to some curve statistical estimation models. Ghosal et al. (2020) employed the linear regression analysis to predict the number of deaths in India due to SARSCoV-2. Ceylan (2020) applied the auto-regressive integrated moving average (ARIMA) model to predict the prevalence of COVID-19 in Italy, Spain, and France.

Ibrahim and Oladipo (2020) in their research forecasted the spread of COVID-19 in Nigeria using Box-Jenkins Modeling Procedure. The ARIMA (1,1,0) model was selected among several ARIMA models based upon the parameter test and Box–Ljung test. A ten-day forecast was also made from the model, which shows a steep upward trend of the spread of the COVID-19 in Nigeria within the selected time frame of February 27th to april 26. In addition, Adeyeri et al (2020) added their voice to the existing knowledge of COVID–19 dynamism in Nigeria by modeling and forecasting the incidence, cumulative incidence and cumulate death using the attack rate, maximum likelihood, exponential growth, Markov chain Monte Carlo, time-dependent and the sequential Bayesian approaches of estimating the COVID–19 reproduction number. Subsequently, these variables are projected for the future. The uncertainty associated with the reproduction numbers as well as the

reproduction ratio sensitivity to the estimation time-period is quantified whilst the correlation between COVID–19 incidence and some meteorological variables are examined. Their findings showed a continuous increase in the incidence, cumulative confirmed cases and cumulative death. Debesh et al. (2020) in their study forecasted COVID-19 confirmed cases in different countries such as China, Italy, South Korea, Iran and Thailand using ARIMA models. Their findings revealed that Mainland China and Thailand were successful in haltering COVID-19 epidemic. Investigating their protocol in this control like quarantine should be in the first line of other countries' program.

Though different times series techniques have been deployed in modelling death due to COViD-19 but the accuracy of these techniques have not been tested hence need for this research. As a result of this, the paper intends to test the accuracy of ARIMA and ARIMAX in modelling number of death with number of confirmed cases as exogenous variable.

2.0 Methodology

The study used secondary data extracted on weekly basis for 43 weeks from NCDC website. In addition, generalized Box Jenkins ARIMA modeling technique coupled with the addition of exogenous variable (ARIMAX) was also adopted to model number of death cases taking number of confirmed cases as exogenous variable. These two techniques were then subjected to performance measures in order to know the model that best predict the data.

2.0.1 Autoregressive Integrated Moving Average Models

An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data under stationarity condition either to better understand the data or to predict future points in the series.

Generally, when stationarity is achieved by differencing then we have ARIMA (p,d,q) model where p, d, and q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model respectively. ARIMA models form an important part of the Box-Jenkins approach to time series modelling. The ARIMA model is as presented in equation 1 below:

$$A_t = \alpha + \beta_1 A_{t-1} + A\beta_{2_{t-2}} + \ldots + \beta_p A_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q} \quad (1)$$

The generation of time series forecasts using ARIMA is better if no outlying data occur. Based on the behavior of the time series, outliers could have potential impact on the estimates on the model parameters. The outlying data could be a pointer to significant events or exceptions and provide useful information to the management. This calls for external variables, which could deliver meaningful answers to the outlying data. As an alternative of modeling a time series A_t with only a combination of past values, A_t can be explained by external variable(s) (regressors).

The ARIMAX model is a ARIMA model with external variables, called ARIMAX $(p, d, q)(X)$, where X is the vector of external variables.

An autoregressive model with exogenous variables (ARX) can be expressed as:

$$A_t = \phi(\mathrm{L})\, A_t + \beta X_t + \epsilon_t \quad (2)$$

The moving average model with exogenous variables can also be expressed as:

$$A_t = \beta X_t + \theta(\mathrm{L})\, \epsilon_t \quad (3)$$

Combining the two, gives:

$$\phi(\mathrm{L})\, A_t = \beta X_t + \theta(\mathrm{L})\, \epsilon_t \quad (4)$$

3.0 Result and Discussion

Table 1: Descriptive Statistics of COVID-19 Confirmed Cases and Deaths

Descriptive Stat.	**Confirmed Cases**	**Deaths**
Minimum	1	0
1st Quarter	612	10
Median	1546	28
Mean	1442	26
3rd Quarter	2154	42
Maximum	3079	56

Source: R-Studio Output

Descriptive statistics of confirned cases and deaths due to covid-19 on weekly basis can be evidenced from table 1. Analysis indicated that the maximum number of confirmed cases were 3079 cases were recorded in a week with an average of 1442 infected people with average death of 26 people.

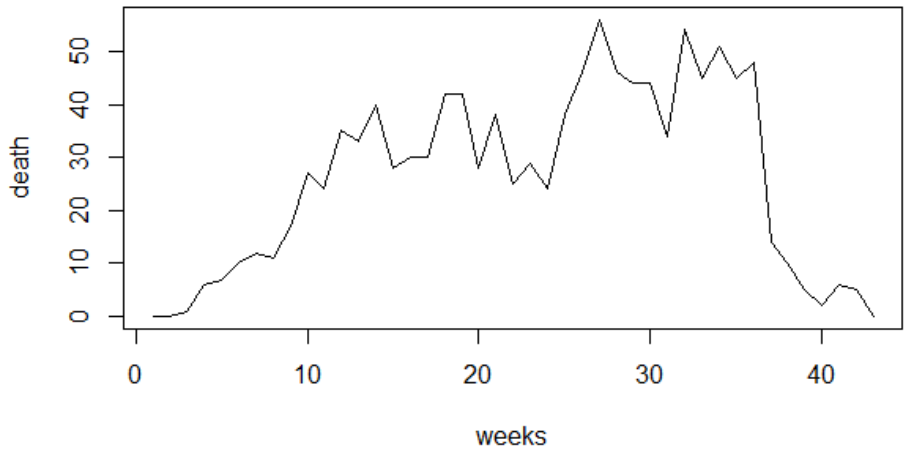

Figure 1: Cumulative *Weekly death cases of COVID-19 Pandemic in Nigeria*

The time series plot of figure 1 shows the number of death cases due to Covid-19 infection. The series indicates non-stationarity with no element of seasonality which can also be evidenced from the ACF and PACF plots of figure 2 and 3.

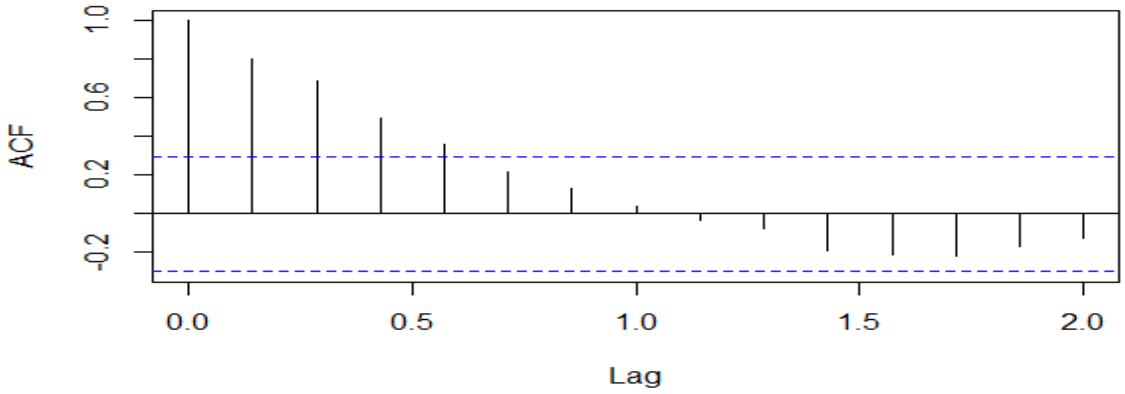

Figure 2: ACF plot of Weekly COVID-19 Death Cases in Nigeria

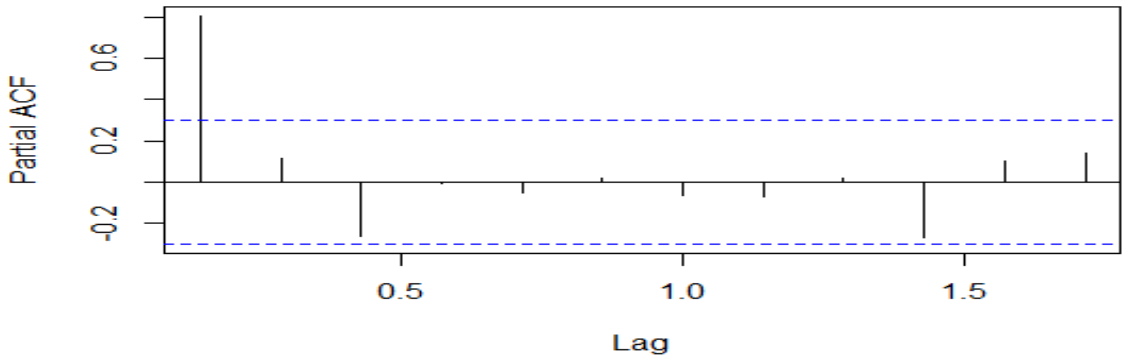

Figure 3: PACF plot of Weekly COVID-19 Death Cases in Nigeria

A significant spike of the ACF and PACF plots taking the death cases into consideration indicated that number of death cases was non-random. The non-randomness of this death cases might be as a result of coronavirus patients having complications which might result in the low survival rate of the dreaded diseases.

Table 2: Unit Root Test of Stationarity; H_0: The Series has a Unit Root

Variable	**ADF @ Levels**	**p-value**	**ADF @ First Difference**	**P-value**	**Remarks**
Death	-0.90562 [3]	0.9411	-3.9693[3]**	0.02043	**I(1)**

ADF Critical Value at 5% = -2.95;
[8] Indicates that a maximum lag length of 8 was included in the tests.

** indicates significant at 5%
Source: Extracted from R-Studio Output

Table 2 confirmed the presence of unit root in the death cases data as evidenced from the Dickey-Fuller test and associated P-values > 0.05 level of significance. We therefore infer that unit root is present in the series. Calibrating the two series for model building development can only be achieved by differencing since ARIMA modeling cannot achieve white noise with a non-stationary data. First order differencing of this series achieved stationarity and was confirmed as order I(1). The ACF and PACF plots after differencing can be seen in figure 4.6 and 4.7 respectively

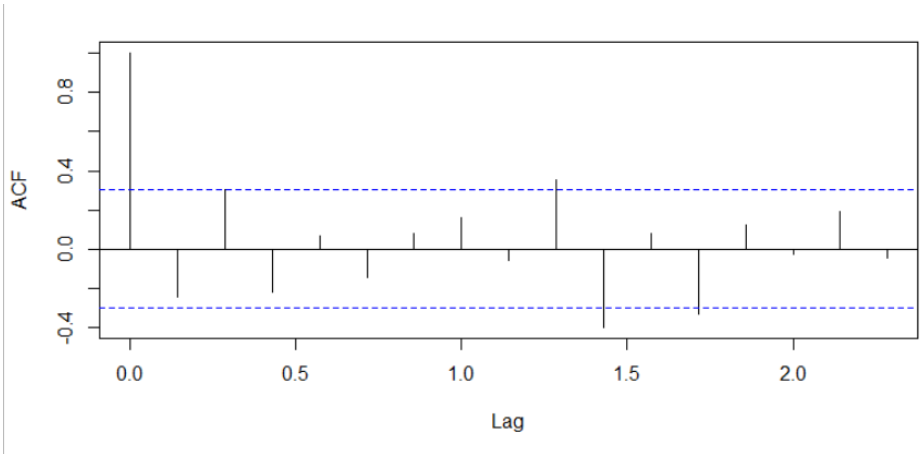

Figure 4: ACF plot of First Differenced Weekly COVID-19 Death Cases in Nigeria

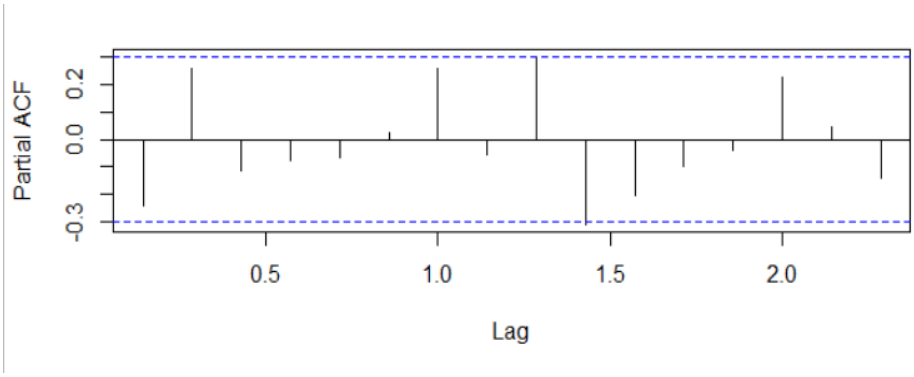

Figure 5: PACF plot of First Differenced Weekly COVID-19 Death Cases in Nigeria

First differenced Correlogram and Partial correlogram study indicates few significant spikes at the respective lags. This can be evidenced that the series is stationary at first difference.

After vigorous ARIMA procedure, the best model is ARIMA (1,1,1) with the mode etimates presented in table 2 below:

Table 3: Best Fitted ARIMA (1,1,1) Model Estimates

Parameters	**Estimates**	**Standard Error**	$\boldsymbol{Z\ value}$	**Pr(>|Z|)**
ϕ_1 AR(1)	-0.2905	0.1556	-1.8666	0.0620*
θ_1 MA(1)	-0.8908	0.0857	-10.3925	0.0000***

**** indicates the significance of the coefficients @1% level*
** indicates the significance of the coefficients @10% level*
Source: R-Studio Output

The model specification for ARIMA (1,1,1) in table 3 is written in form of backshift operator as;

$$(1 - \phi_1 B)A_t = (1 - \theta_1 B)\varepsilon_t \quad (5)$$

Substituting the coefficients, we have;

$$(1 + 0.2905B)A_t = (1 + 0.8908B)\varepsilon_t \quad (6)$$

Table 3 indicate that estimated parameters of AR (1), MA(1) are statistically significant at 1% and 10% level of significance. This implies that Death Cases due to COVID-19 pandemic can well be predicted using ARIMA (1,1,1) model where order of Autoregressive and Moving

Similar result was obtained for ARIMAX model, with the resulted presented in table 4.

Table 4: Fitted ARIMAX Model Estimates

Parameters	**Estimates**	**Standard Error**	$\boldsymbol{Z\ value}$	**Pr(>|Z|)**
ϕ_1 AR(1)	-0.73763315	0.18918713	-3.8990	0.0000***
θ_1 MA(1)	0.50859601	0.22578638	2.2526	0.02429*
Confirmed Cases	0.00072063	0.00331798	0.2172	0.82806

**** indicates the significance of the coefficients @1% level*
** indicates the significance of the coefficients @5% level*
Source: R-Studio Output

The ARIMAX model fitted is written in equation 7 as;

$$(1 + 0.73763315B)A_t = (1 + 0.50859601B)\varepsilon_t + 0.00072063\ (confirmed_{cases}) \quad (7)$$

In a more generalized linear form, we have

$$A_t = +0.5086\varepsilon_{t-1} - 0.73763315X_{t-1} + 0.0007(confirmed_{cases}) \quad (8)$$

Where A_t is the number of deaths due to COVID-19 pandemic on weekly basis.

The inclusion of confirmed COVID-19 cases as an exogenous variable was due to the fact that death due to COVID-19 is dependent on the number of reported cases. As a result of this, the number of confirmed cases was fitted into the ARIMA (1, 1, 1) as an endogenous variable to confirm its level of significance and to also know if the prediction accuracy would be higher than the generalized Box Jenkins model. Although, the order of AR and MA were found to significantly contribute to the ARIMAX model (p-value < 0.5), but the exogenous variable was found to insignificantly contribute to the model but might be adjudged the best fit compared to the traditional ARIMA model fitted.

3.0.1 Diagnostic Check on the fitted ARIMA and ARIMAX models

Table 5: Shapiro-Wilk Test of Residual Normality

Models	Shapiro-Wilk	p-value
ARIMA	0.93238	0.0956
ARIMAX	0.93567	0.0828

Source: R-Studio Output

The Shapiro-Wilk test of normality in table 5 has a test statistic of 0.93238 and 0.93567 for ARIMA and ARIMAX models, with corresponding p-values of 0.0956 and 0.0828 > 0.01 level of

significance where normality of residuals of the best fitted ARIMA (1, 1, 1) and ARIMAX (1,1,1) with predictor variable of number of confirmed cases were not rejected at 1% significance levels. This indicates that the residuals are Normally, Independently and Identically Distributed *(N.I.I.D).*

Table 6: Ljung-Box Portmanteau Test of No serial Correlations

Towns	Chi-Squared	Degree of Freedom	P-value
ARIMA	0.013155	1	0.9087
ARIMAX	0.15676	1	0.6922

Source: R-Studio Output

The small Chi-square statistic and large p-values in the Box test of the fitted ARIMA and ARIMAX residuals (table 6) suggested the failure in rejecting the null hypothesis that all of the autocorrelation functions are zero. In other words, we can conclude that there is no (or almost nil) evidence for non-zero autocorrelations in the residuals of the fitted model. This also indicates that the model has captured the dependence in the series. The AIC and Log-likelihood deal with the fit and parsimony of the model which provides a measure of efficient and parsimonious prediction. In addition, both models fitted can be adjudged to be adequate for predicting the pattern of rate of coronavirus death cases in Nigeria.

Table 7 Predictions Evaluation for COVID-19 Death Cases

Models	MAE	RMSE
ARIMA	6.2152	8.962229
ARIMAX	5.9774	8.538211

Source: R-Studio Output

The best deaths due to COVID-19 pandemic prediction model between ARIMA and ARIMAX model was selected based on its prediction performance using two criteria, namely root mean squared error (RMSE) and mean absolute error (MAE). Table 7 shows ARIMAX model has the

lowest values of RMSE (i.e. 8.538211) and MAE (5.9774) as compared to ARIMA model with RMSE value of 8.962229, and MAE value of 6.2152. This showed ARIMAX model has performed better compared to ARIMA model and therefore, ARIMAX model was selected as the best disease forecasting model in this study.

4.0 Conclusion

This study modelled deaths due to COVID-19 using exogenous variable (number of confirmed cases) in Nigeria using ARIMA and ARIMAX an extension of popularly recognized Box-Jenkins modeling with the aim of identifying the model that will best describe the number of death due to corona virus.

Comparative analysis of the two models fitted indicated that ARIMAX (1,1,1) with confirmed cases as exogenous variable model was found to perform better compared to ARIMA model and therefore, ARIMAX model was selected as the best disease predicting model in this study as shown from the prediction performance. This study corroborates with the work of Ling et al (2019) where they applied ARIMAX model in forecasting weekly cocoa blackpod disease incidence.

5.0 References

Allard, R. (1998). Use of time-series analysis in infectious disease surveillance—*bulletin of the* World Health Organization, *76*(4), 327.

Drexler, J. F., Gloza-Rausch, F., Glende, J., Corman, V. M., Muth, D., Goettsche, M. & Drosten, C. (2010). Genomic Characterization of Severe Acute Respiratory Syndrome-Related Coronavirus in European Bats and Classification of Coronaviruses Based on Partial RNA-Dependent RNA Polymerase Gene Sequences. *Journal of Virology*, *84*(21), 11336–11349. https://doi.org/10.1128/jvi.00650-10.

Ibrahim R.R & Oladipo H.O. (2020): Forecasting the spread of COVID-19 in Nigeria using Box-Jenkins Modeling Procedure. https://doi.org/10.1101/2020.05.05.20091686.

Jiabing, W. (2007). Prediction of Incidence of Notifiable Contagious Diseases by Appalication of Time Series Model. *Journal of Mathematical Medicine*, *1*.

Jin, R., Qiu, H., Zhou, X., Huang, P., Wang, Z., & Wei, J. (2008). Forecasting incidence of intestinal infectious diseases in mainland China with ARIMA model and GM (1, 1) model. *Fudan University Journal of Medical Sciences*, *5*(010).

Jung, S.M., Akhmetzhanov, A.R., Hayashi, K., Linton, N.M., Yang, Y., Yuan, B., Kobayashi, T., Kinoshita, R. & Nishiura, H. (2020). Real-Time Estimation of the Risk of Death from Novel Coronavirus (COVID-19) Infection: Inference Using Exported Cases. Journal of Clinical Medicine, 9(2):523.

Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics, 15(1), 276.

Liu, L., Luan, R. S., Yin, F., Zhu, X. P., & Lü, Q. (2016). Predicting the incidence of hand, foot and mouth disease in Sichuan province, China using the ARIMA model. Epidemiology & Infection, *144*(1), 144–151.

Liu, Q., Liu, X., Jiang, B., & Yang, W. (2011). Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. BMC Infectious Diseases, *11*(1), 218.

Musa, M. I. (2015). Malaria disease distribution in Sudan using time series ARIMA model. International Journal of Public Health Science, *4*(1), 7–16.

Organization, W. H. (2020). Coronavirus disease 2019. *World Health Organization*, https://doi.org/10.1001/jama.2020.2633

Peiris, J. S. M., Lai, S. T., Poon, L. L. M., Guan, Y., Yam, L. Y. C., Lim, W., …, Yuen, K. Y. (2003). Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet, *361*(9366), 1319–1325. https://doi.org/10.1016/S0140-6736(03)13077-2

Sahin, A. R. (2020). 2019 Novel Coronavirus (COVID-19) Outbreak: A Review of the Current Literature. Eurasian Journal of Medical Investigation, *4*(1), 1–7. https://doi.org/10.14744/ejmo.2020.12220

Sato, R. C. (2013). Disease management with ARIMA model in time series. *Einstein (Sao Paulo)*, *11*(1), 128–131.

Seven days in medicine: 11-17 March 2020. (2020). *BMJ (Clinical Research Ed.)*, *368*, 1073. https://doi.org/10.1136/bmj.m1073

Woo, P. C. Y., Huang, Y., Lau, S. K. P., & Yuen, K. Y. (2010). Coronavirus genomics and bioinformatics nalysis. Viruses, 2, 1805–1820. https://doi.org/10.3390/v2081803

Wu, W., Guan, P., Guo, J. Q., & Zhou, B. S. (2008). Comparison of GM (1, 1) gray model and ARIMA model in forecasting the incidence of hemorrhagic fever with renal syndrome. Journal of China Medical University, 37(1), 52–55.

Yin, Y., & Wunderink, R. G. (2018). MERS, SARS and other coronaviruses as causes of pneumonia. *Respirology*, 23, 130–137. https://doi.org/10.1111/resp.13196

Yue, Z., Shengnan, W., & Yuan, L. (2015). Application of ARIMA model on predicting monthly hospital admissions and hospitalization expenses for respiratory diseases. Chinese Journal of Health Statistics, 2, 197–200.

Zaki, A. M., Van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E., & Fouchier, R. A. M. (2012). Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. New England Journal of Medicine, 367(19),