REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1 (2024)
http://www.veterinaria.org

Submitted- 13/April/2024 Reviewed- 07/May/2024 Accepted- 02/June/2024 Published-22/June/2024

A Study Of Pandemic Cases In South Africa Using ARIMA Model

Himanshu Bhatt[1], Manish Karamwar[2*], and Ragesh P.R[3]

[1]Department of Mathematics, University of Delhi, Delhi - 110007, INDIA. Email: du.himanshu@gmail.com
[2*]Department of African Studies, Faculty of Social Sciences, University of Delhi, Delhi -110007, INDIA.
Email: mkaramwar@as.du.ac.in
[3]Department of Zoology, Zakir Husain Delhi College (University of Delhi), JLN Marg, Delhi - 110002, INDIA. Email: rageshpr@zh.du.ac.in

***Corresponding Author**: Manish Karamwar
Department of African Studies, Faculty of Social Sciences, University of Delhi, Delhi -110007, INDIA. Email: mkaramwar@as.du.ac.in

ABSTRACT

Inevitable wildlife habitat destruction, ecological disbalance, and increased human- wildlife interaction have increased the chances of zoonotic diseases spillover in the human population. In the current scenario, an effective and precise forecasting method is needed to control the spread of disease and make appropriate decisions. The present study uses the ARIMA model to forecast COVID-19 cases in South Africa. RMSE, ME, AIC, and BIC measures have been used to select the best- fitted ARIMA model. This model can also be used for other disease forecasting.

Mathematics Subject Classification (2010): 92C60, 92D30, 60G25

Key words and phrases: South Africa, Pandemic, Infection, ARIMA Model, Root Mean Square Error.

Introduction

Anthropocentric strategies of development and negligence to the environment and ecological health paved the way to increased human-wildlife encounters and, thus, the emergence of infectious diseases (EIDs) or zoonotic disease spillover. In the last two decades, the frequency of occurrence of EIDs has increased tremendously. Wildlife animals function as reservoirs of numerous viral pathogens. For instance, Nipah, Zika, influenza, smallpox, measles, and Coronavirus-2 (SARS-CoV-2) viruses have been reported to be transmitted from wildlife to humans. Recently, Coronavirus-2 (SARS-CoV-2) COVID-19 was first re- ported in Wuhan, China, where the live-animals market spread as a pandemic in March 2020 in South Africa. To control the spread of the disease and take preventive measures, precise and accurate forecasting of daily cases is vital. Forecasting is a component of statistical modeling widely used in epidemiology and many other fields. Since ARIMA approach is deemed unsuitable for usage in complex and dynamic contexts, researchers tend to avoid the use of this approach in forecasting COVID-19 cases in South Africa. However, in the current study, an attempt has been made to use ARIMA model to fore- cast the daily COVID-19 cases in South Africa. Box and Jenkins developed the ARIMA model in 1970 to mathematically describe the changes in time series data [3]. After that, using ARIMA model, many disease outbreaks such as pneumonia and influenza [4], Hand–Foot–Mouth Disease (HFMD) [7], and Hemorrhagic Fever with Renal Syndrome (HFRS) [10] have been predicted. Recently, some work has been done in this direction ([5], [8]).

ARIMA Model

ARIMA model stands for "Autoregressive Integrated Moving Average" and is often used for analyzing and forecasting/predicting future events based on the previous time series data. Seasonal and non-seasonal are the two types of ARIMA models. The non-seasonal ARIMA model contains three components as
(i) Auto-regression "$AR(p)$",
(ii) Integrated "$I(d)$",
(iii) Moving Average "$MA(q)$".
The autoregressive model which shows the present value y_t as a linear combination of the
Lagged values of the variables: $y = c + \sum_{i=1}^{p} u_i y'_{t-i} + \epsilon_t$ where u_i for $i = 1,2,\dots,p$ represents the autoregression parameters, p, ϵ_t, c denotes the number of lags, error terms, and constant respectively. Integrated "$I(d)$", which represents to which degree the variable y_t is stationary. In order to enable the time series data to become stationary, Integrated "$I(d)$" depicts the differencing of raw observations (i.e., data values are substituted by the difference between the current and the previous data values) Moving Average "$MA(q)$" expresses the present value of the variable y'_t as a linear combination of the previous error terms i.e. $y'_t = \mu + \epsilon_t + \sum_{i=1}^{q} v_i \epsilon_{t-i}$ where q denotes number of error terms included in the model, μ denotes constant and $v_1, v_2, \dots, v_q$are the parameters of the moving average model.

An ARIMA model can be expressed with the following equation:

$$y'_t = c + \sum_{i=1}^{p} u_i y'_{t-i} + \sum_{i=1}^{q} v_i \epsilon_{t-i} + \epsilon_t \tag{1}$$

where y'_t denote differenced data series,$u_i, (i = 1,2, \dots, p)$ denote the parameters of the autoregression (AR) and $v_i, (i = 1,2, \dots, q)$ denote parameters of the moving average (MA), ϵ_t, c denote the white noise error and constant.
The ARIMA model includes three parameters (p, d, q). The parameters p show the order of the autoregressive part of the model, and parameter d represents that d number of difference transformations are required to vanish the seasonality or trend in data and transform the data into stationary one, i.e., over time making, mean and standard deviation constant. The parameter q represents the order of the moving average part of the model.

Estimating the Parameters through ARIMA Model

Fitting an ARIMA model includes several steps:

(a) Identify whether the time series data is stationary or not. Making data stationary may require some transformation (such as differencing and logging). Using the ADF(Augmented Dickey-Fuller) test [6], KPSS (Kwiatkowski-Phillips-Schmidt- Shin) test, we check the stationarity of the data series and determine the parameter d and using an Auto Correlation function (ACF) and the Partial Auto Correlation function (PACF) we determine p and q.

(b) Estimate the parameters $(c, u_1, u_2, \dots, u_p, v_1, v_2, \dots, v_q)$ using maximum likelihood. Among possible models, the Akaike information criterion (AIC) [1] and Bayesian Information Criterion (BIC) model selection criterion are used for selecting an appropriate model. The model with the lowest AIC and BIC value was selected as the best one.

(c) In the next step, we forecast the data using the final ARIMA (p, d, q) model and validate the forecasted data. The prediction of the model is verified using a number of indicators, including the mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE), which show how the computed values differ from the actual values. The ACF plot determines the presence of autocorrelation between the residual values. Plots depicting the difference between the real and the forecast are used to verify the prediction's correctness.

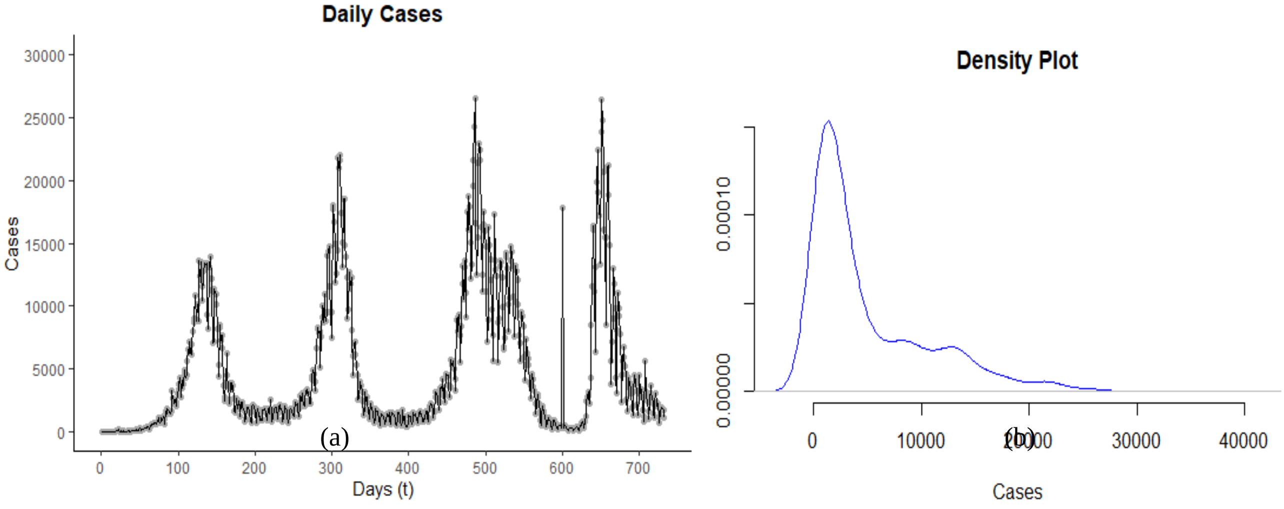
Research Data

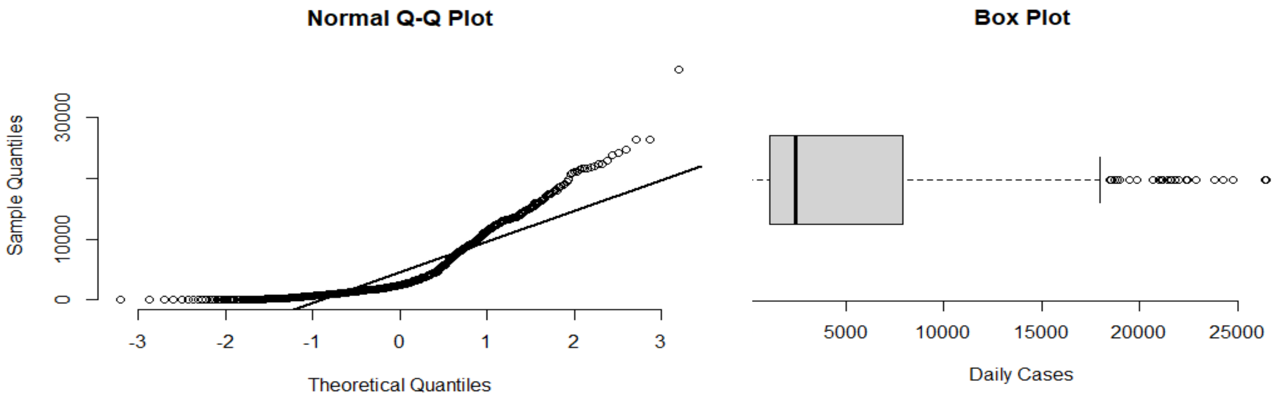
We used the number of daily COVID-19 confirmed cases in South Africa data from Statista [11] from 05 March 2020 to 06 March 2022. For selecting the model, we divide the data into two parts: training data from 05 March 2020 to 02 February 2022 and testing data from 03 February 2022 to 06 March 2022.

Minimum	First Quartile	Median	Third Quartile	Maximum	Mean
0	1102	2432	7907	37875	5033

Table 1: Descriptive Statistics for Covid-19 cases in South Africa

Table (1) shows the number of daily COVID-19 cases in South Africa ranged from 0 to 37875 with an average of 5033 each day from day March 05, 2020, to March 06, 2022, and 2432, the median of the daily Covid-19 cases.





(c) (d)

Figure 1: (a) Daily Covid-19 cases in South Africa, (b) Density Plot, (c) Normal Q-Q Plot and (d) Box-Plot

Figure (1) depicts the plot of the Daily cases, Density plot, Normal Q-Q plot, and Box plot to describe the data. We utilize these plots to check the normality of data. From the density plot and box plot, we conclude that data is skewed towards the right As a Normal Q-Q plot indicates the normality check of the data. Here, a Normal Q-Q plot depicts some deviation from normality. Hence, the data is not normally distributed.

Analysis of Data through ARIMA Model

In this section, we choose an appropriate ARIMA(p,d,q) model using available daily COVID-19 cases in South Africa data. For the ARIMA model, first, we check the stationarity of the daily case data at a 5% First, we test the stationarity of the original time series data, i.e., the daily COVID-19 cases in South Africa data. The ADF test results (Dickey-Fuller $= -2.7804$, Lag order $= 8$, p-value $= 0.248$) show the original data is nonstationary, and the KPSS test depicts that (KPSS Trend $= 0.16339$, Truncation lag parameter $= 6$, p-value $= 0.03551$) original data is stationary. We apply the first difference on the data and then the ADF test for the first differenced data, which depicts stationarity in its mean and variances. Hence, we take the value for the parameter d in the ARIMA (p,d,q) model $d = 1$.R software is used to compute all tests, forecasting the possible candidate for $(p = 5, d = 1, q = 2)$. Through Table (2), we observed the performance of some other ARIMA models.

ARIMA MODEL	ME	RMSE	MAE	MASE	ACF
Auto ARIMA(5,1,2)	10.44061	1926.45	937.7056	0.8147971	0.0118877
ARIMA(7,1,7)	5.226222	1828.712	858.2705	0.7457739	−0.00108665
ARIMA(8,1,7)	5.075643	1816.243	855.2166	0.7431203	−0.00243536
ARIMA(8,1,8)	4.959988	1815.58	862.9408	0.749832	0.00025143
ARIMA(8,1,9)	5.033166	1820.317	847.2287	0.7361794	0.00031343
ARIMA(9,1,8)	5.059093	1815.939	852.3109	0.7405955	−0.00407314

Table 2: ARIMA (p,d,q) model candidates

Table (2) depicts ARIMA(9,1,8) has the smallest mean error (ME) and root mean square error (RMSE) based on RMSE criterion [9] we choose ARIMA(9,1,8) as suitable candidate for daily Covid-19 Cases data in South Africa. For the diagnosis of the results of the selected ARIMA(9,1,8) model, we use the Residual plot, Auto Correlation(ACF) plot, and Partial autocorrelation (PACF) plot (see Figure (2)).

Submitted- 13/April/2024 Reviewed- 07/May/2024 Accepted- 02/June/2024 Published-22/June/2024

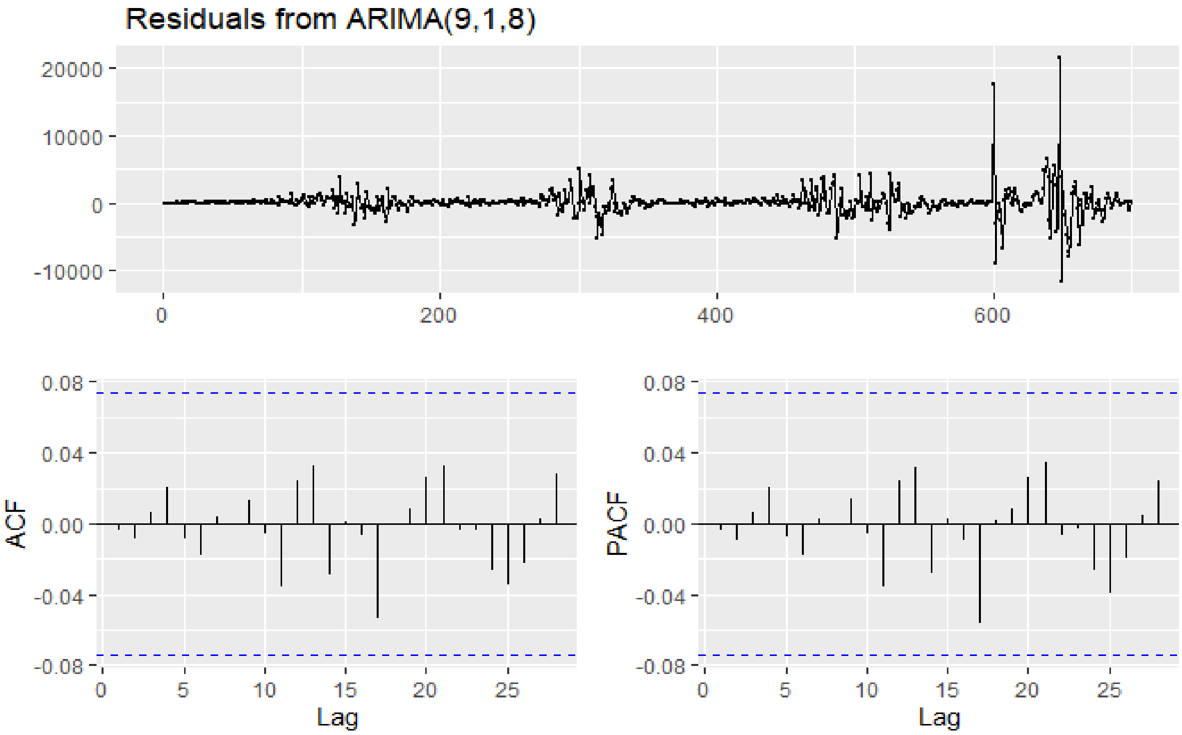


Figure 2: Diagnosis of Residuals from ARIMA(9,1,8)

Model Parameters	Estimate	Standard Error
AR(1)	0.3833	0.9394
AR(2)	0.0295	0.948
AR(3)	−0.2181	0.571
AR(4)	0.0834	0.1846
AR(5)	−0.3579	0.126
AR(6)	0.2536	0.4017
AR(7)	0.4421	0.5405
AR(8)	−0.2549	0.0879
AR(9)	−0.1535	0.2462
MA(1)	−0.8749	0.9403
MA(2)	0.0848	1.4017
MA(3)	0.2462	0.9676
MA(4)	−0.2398	0.3976
MA(5)	0.4256	0.185
MA(6)	−0.3201	0.4967
MA(7)	−0.048	0.6805
MA(8)	0.3247	0.4451

Table 3: Parameter Estimates for ARIMA(9,1,8)

Table (3) shows the estimates of the parameters to describe ARIMA(9,1,8) model fitted using the training data of daily Covid-19 cases from 05 March 2020 to 02 February 2022. Each of the parameter estimates has an absolute value of less than 1 with small standard errors [2].

Submitted- 13/April/2024 Reviewed- 07/May/2024 Accepted- 02/June/2024 Published-22/June/2024

FORECASTING through ARIMA Model

In this section, we forecast the daily COVID-19 cases in South Africa using the ARIMA(9,1,8) model. For the ARIMA(9,1,8) model, training data is used from 05 March 2020 to 02 February 2022; for testing the model, we used data from 03 February to 06 March 2022. Figure (3) depicts the forecast from the ARIMA(9,1,8) model with 80% and 95% confidence intervals, and Figure (4) shows the Forecasted and Actual number of daily Covid-19 cases in South Africa are close to each other.

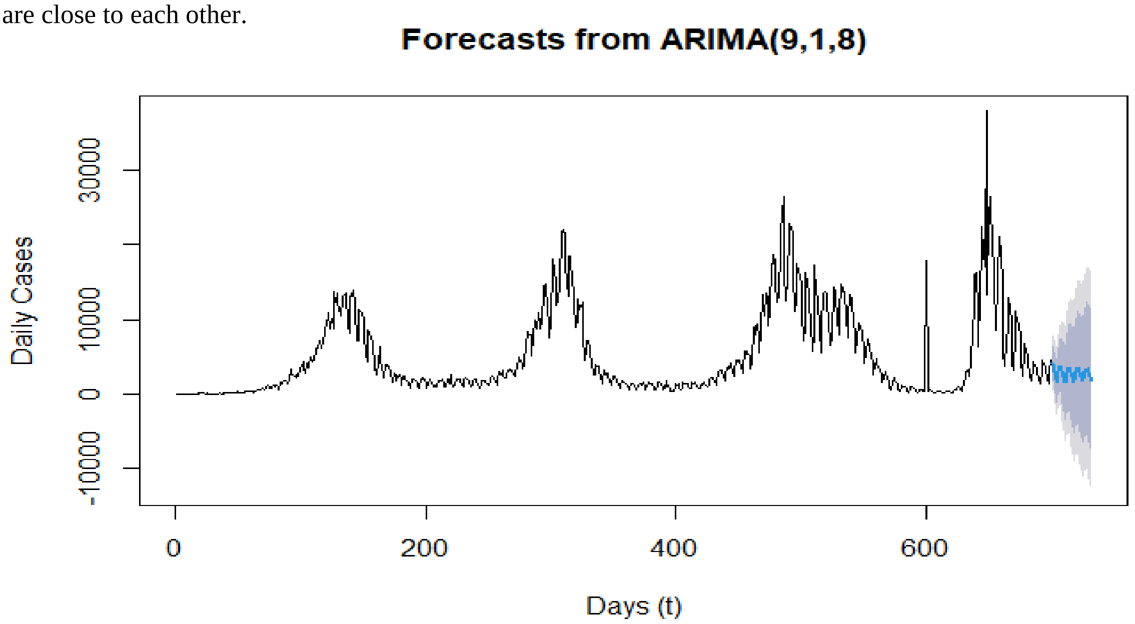


Figure 3: Forecasts from ARIMA(9,1,8) Model with 80% and 95% Confidence Interval

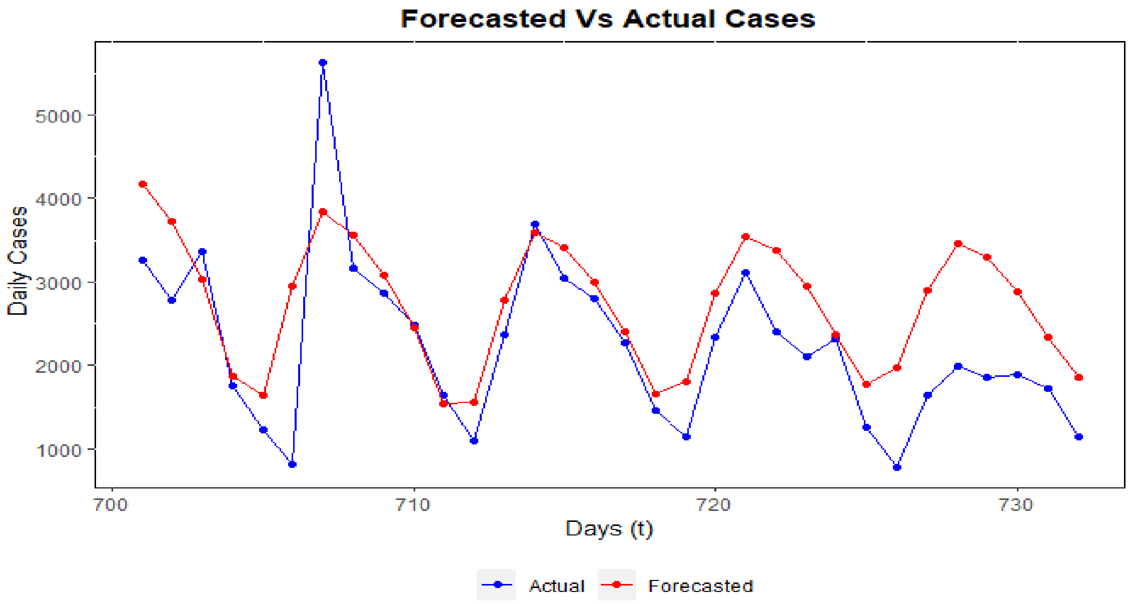


Figure 4: Actual and Forecasted Daily Covid-19 Cases from 03 February 2022 to 06 March 2022

Conclusion

The current study, using the ARIMA model for Covid-19 positive cases data in South Africa as an example study, made an attempt to predict the possible future trends in the occurrence of zoonotic spillovers. Regarding ACF and PACF charts, as well as accuracy measures like RMSE, MAE, ME, AIC, and BIC, the ARIMA model appeared to be best-fitted. The forecasted cases followed the same trend as those reported in South Africa.

The ARIMA model is one of the most effective statistical tools to forecast and predict the spread of pathological diseases in an area. In 2019, private expenditure on health as a share of total health expenditure for South Africa was 40.1. This percentage fell gradually from 62.8% in 2000 to 40.1% in 2019. In 2019, South Africa's health expenditure as a share of GDP was 8.2%. Though it fluctuated substantially in recent years, it tended to increase from 2002 - 2019, ending at 8.2% in 2019. This study enables the statutory bodies/policymakers to plan for the precautionary measures

Submitted- 13/April/2024 Reviewed- 07/May/2024 Accepted- 02/June/2024 Published-22/June/2024

that can be taken to prevent the spread of the diseases.

References

[1] Akaike, H. (1974). *A new look at the statistical model identification*. IEEE transactions on automatic control, 19(6), 716–723.

[2] Alabdulrazzaq, H., Alenezi, M.N., Rawajfih, Y., Alghannam, B.A., Al-Hassan, A.A., & Al-Anzi, F.S. (2021). *On the accuracy of ARIMA based prediction of COVID-19 spread*. Results in Physics, 27, 104509.

[3] Box, G.E., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[4] Choi K., Thacker S.B. (1981). *An evaluation of influenza mortality surveillance, 1962–1979. I. Time series forecasts of expected pneumonia and influenza deaths*. Am J Epidemiol. 113(3), 215—226.

[5] Claris, S. & Peter, N. (2022). *ARIMA model in predicting of covid-19 epidemic for the southern Africa region*. African Journal of Infectious Disease, Published online December 22.

[6] Dickey, D.A., & Fuller, W.A. (1979). *Distribution of the estimators for autoregressive time series with a unit root*. Journal of the American statistical association, 74(366a), 427–431.

[7] 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.

[8] Madan, S., Garg, P., Arora, R. & Singh, D.K. (2022) *Estimating the parameters of covid-19 cases in South Africa*. Biosciences Biotechnology Research Asia, 19(1), 153–162.

[9] Singh, S., Sundram, B.M., Rajendran, K., Law, K. B., Aris, T., Ibrahim, H., ... & Gill, B.S. (2020). *Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models*. Journal of infection in developing countries, 14(9), 971–976.

[10] Zhao, Y., Ge, L., Zhou, Y., Sun, Z., Zheng, E., Wang, X., ... & Cheng, H. (2018). *A new seasonal difference space-time autoregressive integrated moving average (SD- STARIMA) model and spatiotemporal trend prediction analysis for hemorrhagic fever with renal syndrome (HFRS)*. PloS one, 13(11), e0207518.

[11] https://www.statista.com