Predicting South Africa's Daily COVID-19 Cases using ARIMA Forecasting Model: 6 March to 6 July 2020

Shoko Claris (claris.shoko@gmail.com)
Great Zimbabwe University
Chikobvu Delson
University of the Free State

Research Article

Keywords: COVID-19, ARIMA(p,d,q), forecasting, Box-Ljung test, correlogram

Posted Date: May 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-525194/v1

Predicting South Africa's daily COVID-19 cases using ARIMA forecasting model: 6 March to 6 July 2020

Shoko Claris[1] and Chikobvu Delson[2]
***Corresponding author, email**: claris.shoko@gmail.com*

Abstract

Background and Objective: The COVID-19 pandemic caused approximately 11,421,822 laboratory confirmed cases globally with 196,750 confirmed cases in South Africa by the 6th of July 2020. Coronavirus is transmitted from one person to another even before any symptoms appear, thus posing a severe threat to the society as a whole. This study is aimed at coming up with an ARIMA model to predict daily COVID-19 disease cases in South Africa using data from online sources.

Materials and Methods: The study used online data on daily COVID-19 reported cases in South Africa (SA) recorded from 6 March 2020 to the 6th of July 2020. Time series analysis is used to investigate the trend in the daily COVID-19 disease cases leading to the Auto-Regressive Integrated Moving Average (ARIMA) model.

Results: The time plot of the series suggests the need for differencing of the data up to the second-order to achieve a stationary time series. The best candidate model was an ARIMA(7,2,0). Residuals for the selected model are non-correlated and normally distributed with mean zero with a constant variance as expected in a good model. The fitted model predicted a continuous increase in the daily COVID-19 disease cases for the next 20 days ahead to day 143 with slight falls at a few time points.

Conclusion: The results showed that ARIMA models can be applied to COVID-19 patterns in South Afriva. The model forecasted a continuous increase in the daily COVID-19 cases in South Africa. These results are important for public health planning in order combat the pandemic.

Keywords: COVID-19, ARIMA(p,d,q), forecasting, Box-Ljung test, correlogram

INTRODUCTION

The pandemic of the coronavirus disease (COVID-19 disease) started in China (Wuhan, Hubei province) and the first cases were officially recorded on the 31st of December 2019[1]. The number of COVID-19 disease cases accelerated at an alarming rate in China and on the 30th of January, it was declared a Public Health Emergency of International concern. This was the highest recent level in the World Health Organisation's (WHO) emergency response to infectious diseases[2].

Coronaviruses, a large family of viruses, can cause illnesses that range from the common colds to much more severe illnesses like SARS, Middle East respiratory syndrome, and COVID-19[3]. Signs of the COVID-19 disease may include fever, cough, shortness of breath and general breathing difficulties, organ failure, and even death. Chinese health authorities stated that coronavirus is likely to be transmitted from one person to another even before any symptoms appear (spread during the incubation period), making prevention and control difficult. This poses a severe threat to society as a whole. The pandemic caused approximately 11,421,822 laboratory-confirmed cases globally with 196,750 confirmed cases by the 6th of July 2020 in South Africa[4].

Various mathematical and statistical models have been proposed to predict the spread of the COVID-19 disease. These models include the SEIR models[5], and the autoregressive moving average models[6,7]. The autoregressive integrated moving average (ARIMA) model can help to timely analyse and predict the changes in the COVID-19 disease, and provide dynamic information to relevant departments[8]. The ARIMA model is good at forecasting linear time series[9].

Autoregressive integrated moving average (ARIMA) models have been used to predict future development trends of incidence and prevalence in epidemiological data. Benvenuto et al. used an ARIMA model on the Johns Hopkins epidemiological data to predict the epidemiological trends[10]. From their analysis, an ARIMA (1,0,4) and ARIMA (1,0,3) were selected as the best models to determine the early prevalence and incidence of the COVID-19 disease, respectively.

In their research, Duan and Zhang introduced the ARIMA model to analyse daily new COVID-19 disease data sets from Japan and South Korea[8]. They selected ARIMA (6,1,7) and ARIMA (2,1,3) to predict 7 days in advance for Japan and South Korea, respectively.

The objective of this study was to come up with an optimal ARIMA model that best described the trend in daily COVID-19 cases in South Africa. The selected model was used to forecast and predict

20-days ahead to assist the South African government and policymakers to come up with ways to combat the pandemic.

MATERIALS AND METHODS

Data collection:

The data used was obtained online Geographic Distribution of the COVID-19 disease cases[11]. This data is updated daily. The daily COVID-19 disease cases for South Africa were extracted and the data collected from the 6th of March, 2020 to the 6th of July 2020 was used for building the models.

The Time Series Process

A time-series process is a set of random variables $\{X_t, t \in T\}$, where T is the set of times at which the process was observed. The assumption is that each random variable X_t is distributed according to some univariate distribution function F_t. It also considers that time intervals are equidistant and for the real-valued random variables X_t to allow for enumeration of the set of times for $T = \{1, 2, 3, \ldots\}$[12].

Non-Seasonal Autoregressive Integrated Moving Average (ARIMA) models

The growth of daily COVID-19 disease cases for South Africa can be captured like other series, by an integrated model such as the ARIMA[13]. ARIMA models are aimed at describing series which exhibit a trend that can be removed by differencing. The differenced series can be described by an ARMA (p,q) model. If the order of differencing is d, the definition of an ARIMA (p,d,q) process arises naturally. Autoregressive moving average (ARMA) models are given by the following formula[14]

$$y_t = c + \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$$

c is a constant, $\emptyset_i$ are the autoregressive component coefficients and θ_i are the moving average component coefficients. y_{t-i} are lagged values of the observed for $i = 1,2, \ldots . p$. e_{t-j} are lagged error terms for $j = 1,2, \ldots . q$.

In this equation, predictors include lagged values of y_t and lagged error terms, e_t. An ARIMA model is the original model before differencing to the ARMA model. Thus, $(1 - B)^d y_t$, where B is the backshift operator, follows an ARMA model. ARIMA(p,d,q) represents an ARIMA model such that p is the order of the autoregressive component, d is the degree of the differencing involved, and q is the order of the moving average component. Therefore, an ARIMA(1,1,1) is given as:

$$(1 - \emptyset_1 B)(1 - B)y_t = c + (1 + \theta_1 B)e_t$$

B is the backshift operator defined as $By_t = y_{t-1}$.

The ARIMA(1,1,1) can be written out as:

$$y_t = c + y_{t-1} + \emptyset_1 y_{t-1} - \emptyset_1 y_{t-2} + \theta_1 e_{t-1} + e_t$$

Selection of the best ARIMA(p,d,q) model for the data

Goodness-of- fit for the fitted ARIMA(p,d,q) models is done using the Likelihood ratio test (LRT), Akaike Information Criteria (AIC)[15], Mean Absolute Percent Error (MAPE), and Bayesian Information Criteria (BIC)[13]. These methods are defined below:

1. MLE is used to estimate the parameters of the model, that is, $c, \emptyset_1, \ldots, \emptyset_p, \theta_1, \ldots, \theta_q$. The MLE is obtained by minimising $\sum_{t=1}^{T} e_t^2$
2. AIC= $-2\log(L) + 2(p + q + k + 1)$, where L is the likelihood of the data, $k = 1$ if $c \neq 0$ and $k = 0$ if $c = 0$.
3. BIC= $AIC + \log(T)(p + q + k - 1)$
4. Ljung-Box Test: It tests the null hypothesis that several autocorrelation coefficients are simultaneously equal to zero. Or it evaluates whether there is any significant autocorrelation in a series. The test statistic is:

   $$Q(h) = n \cdot (n+2) \cdot \sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{n-k}$$

   where n is the length of the time series, $\hat{\rho}_k^2$ are the sample autocorrelation coefficients at lag k and h is the lag up to which the test is performed. The test statistic asymptotically follows a χ^2 distribution with h degrees of freedom.

Results

In this section, a step-by-step procedure to come up with the optimal ARIMA(p,d,q) for the daily COVID-19 recorded cases for South Africa from the 6th of March 2020 (the day when the first case was recorded) to the 6th of July 2020, thus giving 123 days post identification of the first COVID-19 case. To check for stationarity of the time series data, a time series plot for the observed daily COVID-19 cases is drawn and shown in Fig. 1.

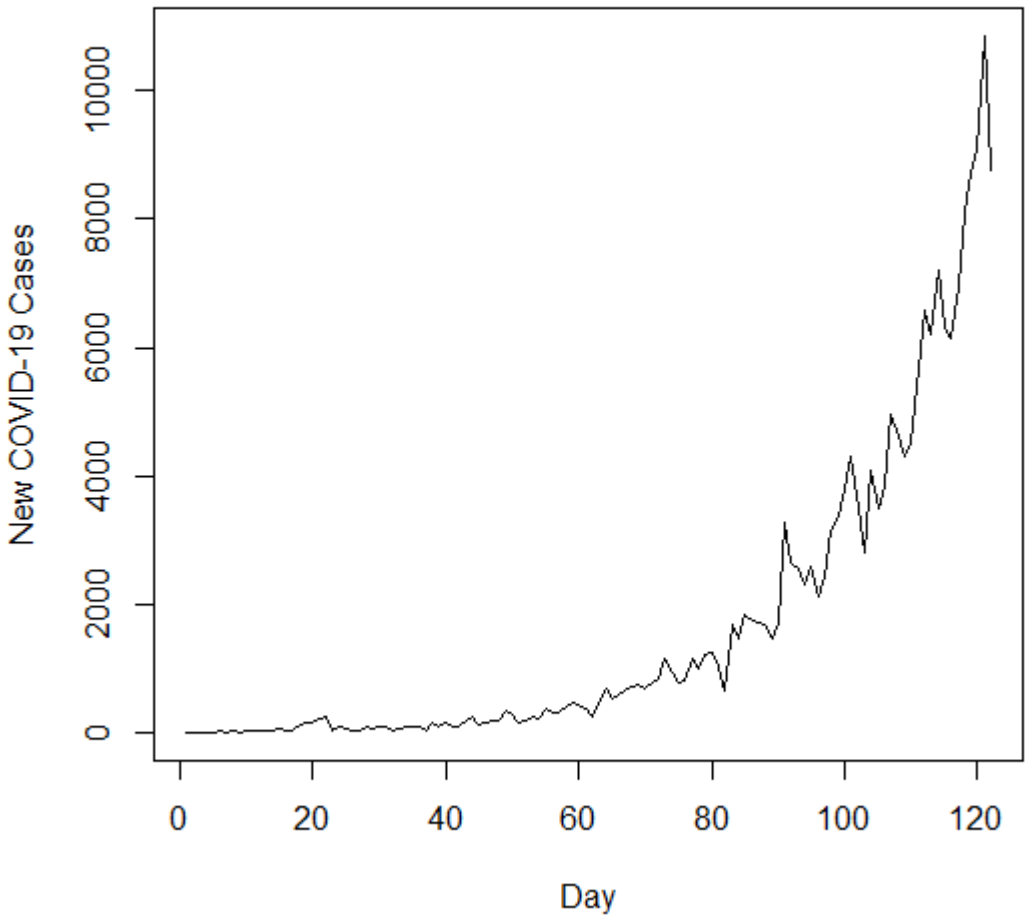

Figure 1: Trajectory of daily COVID-19 cases in South Africa from 6 March to 6 July, 2020

The time series for the new COVID-19 cases in South Africa show an exponential increase in the number of cases. The Dickey-Fuller = 2.6456, Lag order = 4, p-value = 0.99 confirms that the recorded data is not stationary. The important finding is the persistent increase in cases.

The homogeneity check between successive terms in the time series data is done by plotting a lagged-scatter plot for the observed daily COVID-19 cases for South Africa. The lagged scatter plot is shown in Fig. 2.

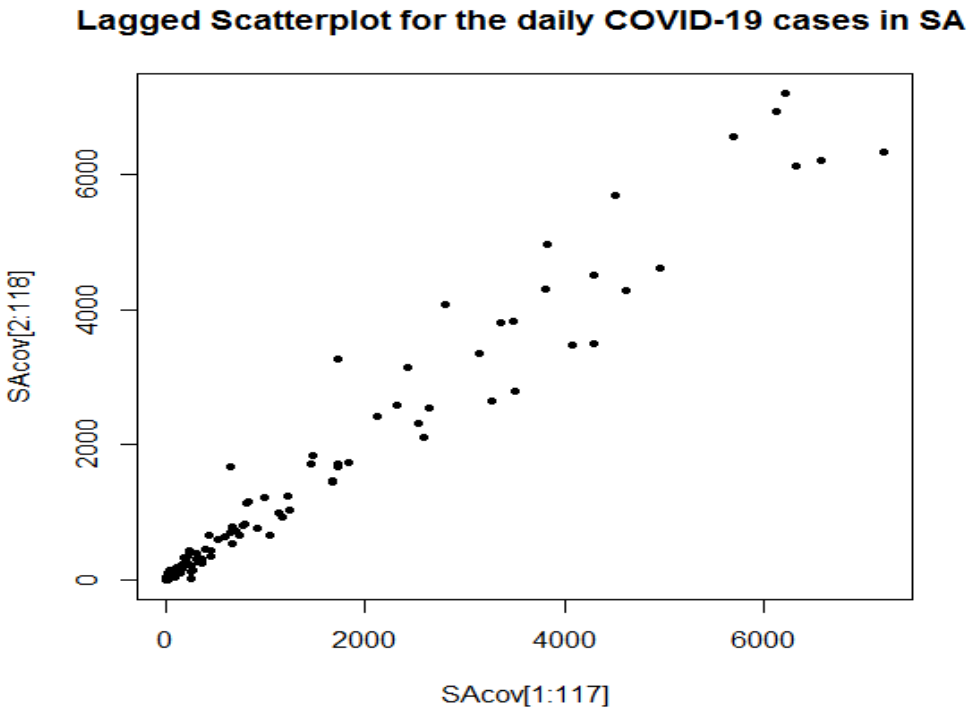

Figure 2: Lagged Scatter plot for the daily recorded cases for South Africa

Figure 2 shows positive correlation between successive measurements. This clearly shows that if the previous observation was above or below the mean, the next one is more likely to be on the same side. The Pearson correlation coefficient turns out to be 0.95, thus a very strong linear relationship. The square of the correlation coefficient, $0.95^2 = 0.90$, is the percentage of variability explained by the linear association between y_t (represented by SAcov[2:118]) and its respective predecessor. Thus, in this case, y_{t-1} (represented by SAcov[1:117]) explains roughly 90% of the variability observed in y_t.

Diagnostic checking

The time series ARIMA model that can be used to predict COVID-19 daily cases for South Africa can only be built once the series is stationary. This is done by taking the n^{th}-order difference (d=n) until the series becomes stationary and then test for stationarity (unit root problem). The test for stationarity is done using the Augumented Dickey-Fuller (ADF) test, under the hypotheses, H_0: Time series data is non-stationary (presence of a unit root) and the alternative, H_a : Time series data is stationary.

The results from the first-difference suggests an ARIMA(7,1,0). However, further investigation on the behaviour of the residuals show that the residual PACF violets the White noise assumption. This was further confirmed by the Box-Ljunk test: X-squre = 22.679, degree of freedom=7, p-value = 0.001939 (close to zero) indicates the need for further differencing.

Results from the second-difference proposes an ARIMA(7,2,0) based on the BIC. The Box-Ljunk test: X-square = 7.2104, df = 7, p-value = 0.4073, shows little evidence of non-zero autocorrelation in the residuals. Thus, the ARIMA (7,2,0) is the best model for predicting and forecasting daily COVID-19 cases in the next 20 days. The estimated parameters for the selected ARIMA (7,2,0) model are shown in Table 1.

Table 1: Estimated parameters for the selected model

Parameters	ar1	ar2	ar3	ar4	ar5	ar6	ar7
Estimates	-1.3936	-1.4631	-1.3169	-1.3258	-1.3807	-1.0775	-0.3554
Standard Er.	0.0909	0.1316	0.1471	0.1450	0.1400	0.1305	0.0998
|z|-value	15.331	11.118	8.952	9.143	9.862	8.257	3.561
Sig. at 5%	***	***	***	***	***	***	**
sigma^2 estimated as 134027: log likelihood=-878.15 AIC=1772.3 ; AICc=1773.59; BIC=1794.6							

Table 1 shows the parameters for the optimal ARIMA model (ARIMA(7,2,0)) for daily COVID-19 cases in South Africa. The selected model tells us of the need to take into account the COVID-19 cases at 7 lags (a week's worth of information) from a given time point t. It also tells us that the time series is not stationary, so we need to take a second order difference. The z-values for the estimated parameters are all greater than 1.96 in absolute value. Thus, all coefficients are significant at 5% level. Hence, the selected model was used to forecast for 20-days in advance in the next subsection.

Forecasting using the selected ARIMA(7,2,0) model

The fitted ARIMA(7,2,0) model, was used to forecast for the future values of our time series of COVID-19 daily cases for the next 20 days (from 07/07/2020 to 26/07/2020) with 80% and 95% (low and high) prediction intervals.

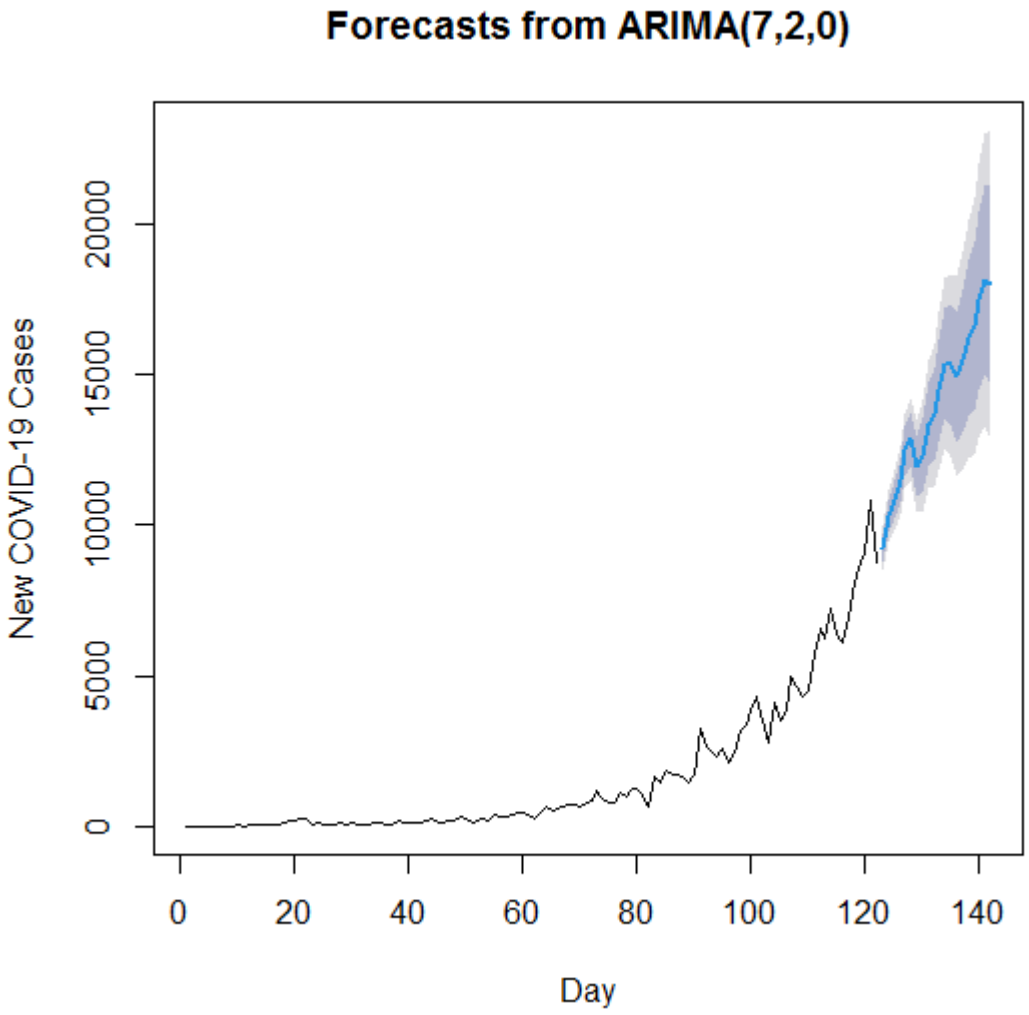

Figure 3: Forecasting plot of trajectory of daily COVID-19 cases in South Africa

Figure 3 gives a clear picture of the trend forecasted time series for the daily COVID-19 cases in South Africa. In Figure 3, the two shaded zones of forecast represent the 80% and 95% (lower and upper side) projection of prediction intervals. The time series forecasts a continued daily increase in the COVID-19 cases for South Africa with peaks after an average of 7 days. Thus, the predicted daily COVID-19 cases in South Africa are estimated to increase from 9,062 (95% CI: 8,364.001; 9,761.016) on 7 July 2020 to approximately 17,446 on 26 July 2020 (95% CI: 14,309.461; 20,582.954).

DISCUSSION

The study employed ARIMA models to forecast ahead for 20-days, the daily COVID-19 cases in South Africa. The data used was obtained from online daily COVID-19 disease cases as updated by WHO. The choice of the ARIMA(p,d,q) model was influenced by the fact that ARIMA models can handle predictions when a time series data shows either an increasing or decreasing trend and autocorrelations between the successive values of the time series[17,18]. The time series plot for the COVID-19 cases in South Africa from the 6th of March to the 6th of July showed a persistent increase in the number of daily cases. The data showed a very positive correlation (Pearson's correlation coefficient = 0.95). The square of the correlation coefficient showed that 90% of the variability observed in any day is explained by its previous day. Thus, the previous day's number of cases is so relevant to the current number of cases, and will in most cases be more. Stationarity of the time series was achieved by differencing and then performing some diagnostic checks on the candidate models. Diagnostic checks were based on residuals plots, AICs, and BICs of the selected models. The Box-L-jung test was also performed on candidate models. This lead to the selection of the ARIMA(7,2,0). After the second difference of the time series, the residuals plot for the selected model inferred that the standard errors are now constant in variance and mean over time, although some slight variations were shown towards the end of the time series. The absolute values of the estimated auto-regressive coefficients were all far below 7, which is a further confirmation that the series was now stationary.

The selected model reveals that a week's information on number of cases is needed to be able to forecast into the future. It also tells us that taking a second order difference is sufficient for the time-series to be stationary.

The successive residuals (forecast errors) for the selected model where statistically tested and are not correlated. The residuals seem to be somewhat normally distributed with mean zero and constant variance. This lead to the conclusion that the ARIMA(7,2,0) provide adequate predictive model for the COVID-19 cases in South Africa over the selected period. This shows that a second-order difference of the time series was sufficient to make the series stationary. This collaborates with results obtained by Tandon et al.[19] and Kufel[20] who also suggested second order difference of the time series models. A study by Chellai et al. showed that the optimal ARIMA model was the ARIMA(7,2,0) for the daily COVID-19 cases in South Africa up to the 19th of March[18]. In this case, the trend in COVID-19 disease cases in South Africa was not well pronounced since it was done during the early stages of the pandemic.

The selected model in this study was used to forecast 20-days ahead, the daily COVID-19 disease cases for South Africa. The ARIMA(7,2,0) predicted a continuous increase in the daily COVID-19

disease cases from day 123 to day 128, a fall on day 129, and an increase in subsequent days up to day 143. In general, the selected model predicted a continuous increase in daily COVID-19 cases from approximately 9,062 on 6 July 2020 to approximately 17,446 on 25 July 2020. Although a studies carried out in India proposed an ARIMA(4,1,1) model, the study also forecasted an increasing trend in the COVID-19 cases from the 30th of June to the 19th of July[21,22]. Thus, the South African government needs to plan for an increase in hospitalisations and death from Covid-19 in this short term period ahead.

However, the data provided is likely to be prone to error due to economic challenges and cost of testing. This might lead to underestimation of the daily COVID-19 cases.

Conclusion

COVID-19 is a challenge, not only to South Africa, but to the world at large. The results indicated that ARIMA models can be applied to forecast COVID-19 cases in South Africa. The predicted persistent rapid increase in the daily recorded cases is a cause of concern.

Significance statement

This study predicted the trend in daily COVID-19 disease cases for public awareness. This study will help the government and policy makers to plan ahead in order to combat the pandemic.

Declarations

Ethics approval and consent to participate

Not Applicable. Data is available online

Consent for publication

Not applicable

Availability of data and materials

The data used was obtained online Geographic Distribution of the COVID-19 disease cases (https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-COVID-19-cases-worldwide). This data is updated daily.

Competing interests

The authors do not have any competing interests

Funding

Not applicable

Authors' contributions

Claris Shoko devised the initial idea and drafted the first manuscript. Delson Chikobvu finalised and proofread the article. Claris Shoko and Delson Chikobvu contributed to the analysis and interpretation of the data. Both authors participated in critical revision of the manuscript drafts and approved the final version

Acknowledgements

We thank the participants of the study

Authors' information (optional)

Claris Shoko: Email: claris.shoko@gmail.com. Department of Mathematics and Computer Sciences, Faculty of Natural Sciences, Great Zimbabwe University, PO Box 1235, Masvingo, Zimbabwe.

Delson Chikobvu: Email: Chikobvu@ufs.ac.za. Department of Mathematical Statistics and Actuarial Sciences, University of the Free State, P.O. Box 339, 9300 Bloemfontein, South Africa.

If any of the sections are not relevant to your manuscript, please include the heading and write

References

1 WHO. Coronavirus disease 2019 (COVID-19) Situation Report – 31. February 2020. Accessed from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200220.

2. WHO, Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV), January 2020, Accessed from: https://www.who.int/news-room/detail/30-01-2020

3. Center for Health Security. Coronaviruses: SARS, MERS, and 2019-nCoV Updated April 14, 2020. Johns Hopkins. Accessed from: https://www.centerforhealthsecurity.org/resources/fact-sheets/pdfs/coronaviruses.pdf

4. Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina and Joe Hasell (2020) - "Coronavirus Pandemic (COVID-19)". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/coronavirus'

5. Fang Y., Nie Y., Penny M. transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis. J Med Virol. 2020; 92: 645-659. DOI: 10.1002/jmv.25750.

6. Panda M. Application of ARIMA and Holt_Winters forecasting model to predict the spreading of COVID-19 for India and its states. 2020 Accessed from: https://doi.org/10.1101/2020.07.14.20153908.

7. Chakraborty T., Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. Chaos, Solitons and Fractals 135 (2020) 109850. https://doi.org/10.1016/j.chaos.2020.109850.
8. Duan X., and Zhang X. ARIMA modelling and forecasting of irregularly patterned COVID-19 outbreaks using Japanese and South Korean data. Data in brief 31 (2020) 105779. https://doi.org//10.1016/j.dib.2020.105779.
9. Box GE , Jenkins GM , Reinsel GC , Ljung GM . Time series analysis: forecasting and control. John Wiley & Sons; 2016. https://doi.org/10.1111/jtsa.12194 .
10. Benvenuto D., Giovanetti M.,Vassallo L., Angeletti S., and Ciccozzi M. Application of the ARIMA model on the COVID-19 epidemic dataset. Data in brief 29 (2020) 105340. http://doi.org/10.1016/j.dib.2020.105340.
11. European Centre for Disease Prevention and Control (https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-COVID-19-cases-worldwide)
12. Adhikari, R. and Agrawal, R. (2013) An introductory study on time series modeling and forecasting. Lambert Academic Publishing. Copyrght 2013. arXiv preprint arXiv:1302.6613. https://arxiv.org/ftp/arxiv/papers/1302/1302.6613.pdf.
13. Chatfield C. The analysis of time series: an introduction. Chapman and Hall/CRC; 2016.
14. Zhang G.P. Time Series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50(2003);159-175. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.3756&rep=rep1&type=pdf.
15. Akaike H. (1974) A New Look at the Statistical Model Identification. In: Parzen E., Tanabe K., Kitagawa G. (eds) Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY. https://doi.org/10.1007/978-1-4612-1694-0_16
16. Schwarz, Gideon E. (1978), Estimating the dimension of a model, Annals of Statistics, 6 (2): 461-464, doi:10.1214/aos/1176344136
17. Adebiyi, A.A.; Adewumi, A.O.; Ayo, C.K., (2014). Comparison of ARIMA and artificial neural networks models for stock price prediction. J. Appl. Math., Article ID 614342 (**7 pages**). https://doi.org/10.1155/2014/614342.
18. Chellai F., Ahmed H., and Pradeep M. COVID-19 Statistics, strange trend and forecasting of total cases in the most infected African countries: An ARIMA and Fuzzy time series approach. ResearchGate, Preprint-March 2020. DOI:10.13140/RG2.2.34158.97603.
19. Kufel, T. ARIMA-based forecasting of the dynamics of confirmed COVID-19 cases for selected European countries. Equilibrium. Quarterly Journal of Economics and Economic Policy, 15(2), 181-204. Doi: 10.24136/eq.2020.009.

20. Tandon A., Ranjan P., Chakraborty T., and Suhag V. Coronavirus (COVID-19): ARIMA based time series analysis to forecast near future, 2020. **https://arxiv.org/abs/2004.07859**

21. Marbaniang S.T. Forecasting the prevalence of COVID-19 in Maharashtra, Delhi, Kerala, and India using an ARIMA model. Research Square, Preprint. DOI:https://doi.org/1021203/rs.3.rs-34555/v1.

Figures

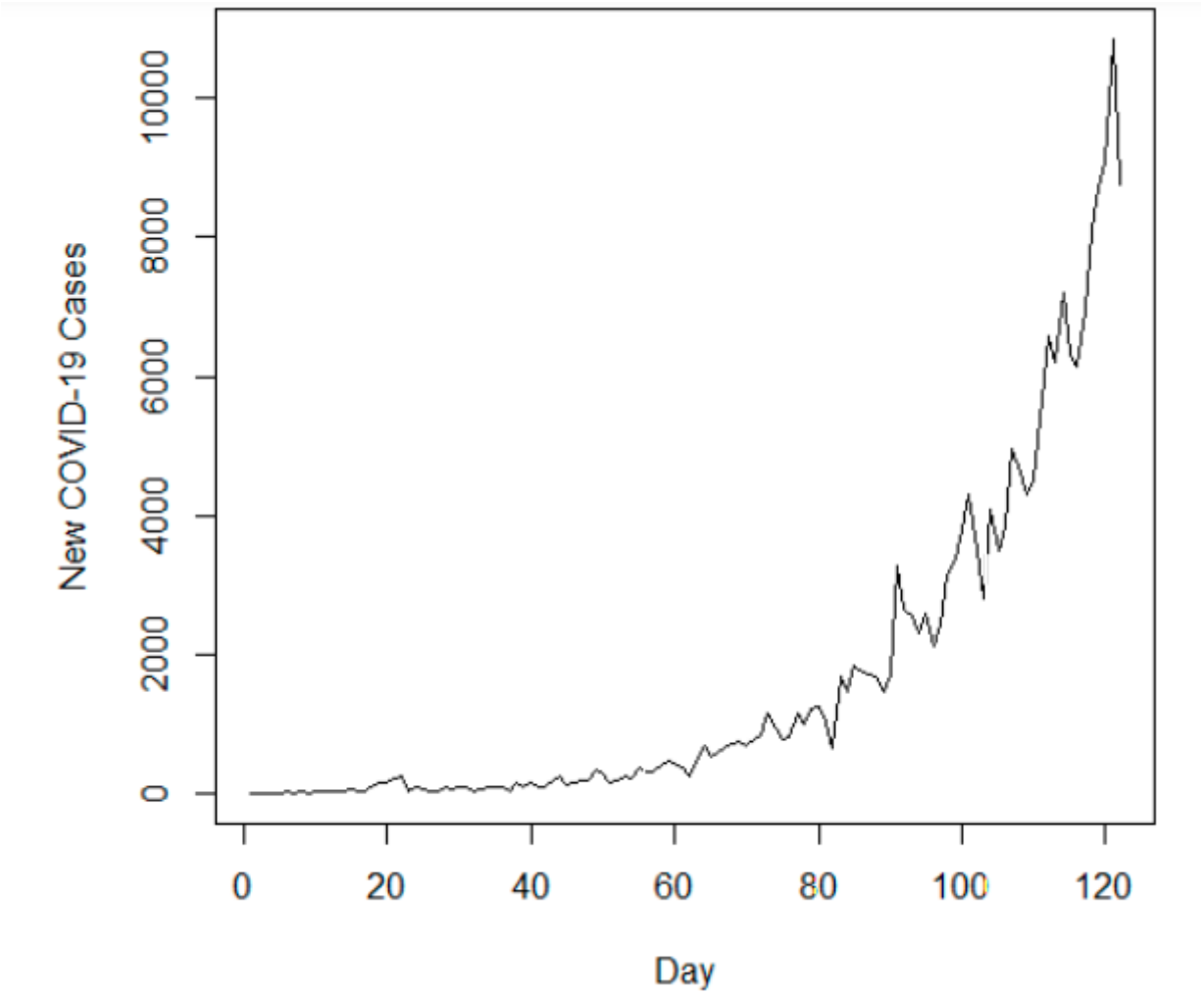

Figure 1

Trajectory of daily COVID-19 cases in South Africa from 6 March to 6 July, 2020

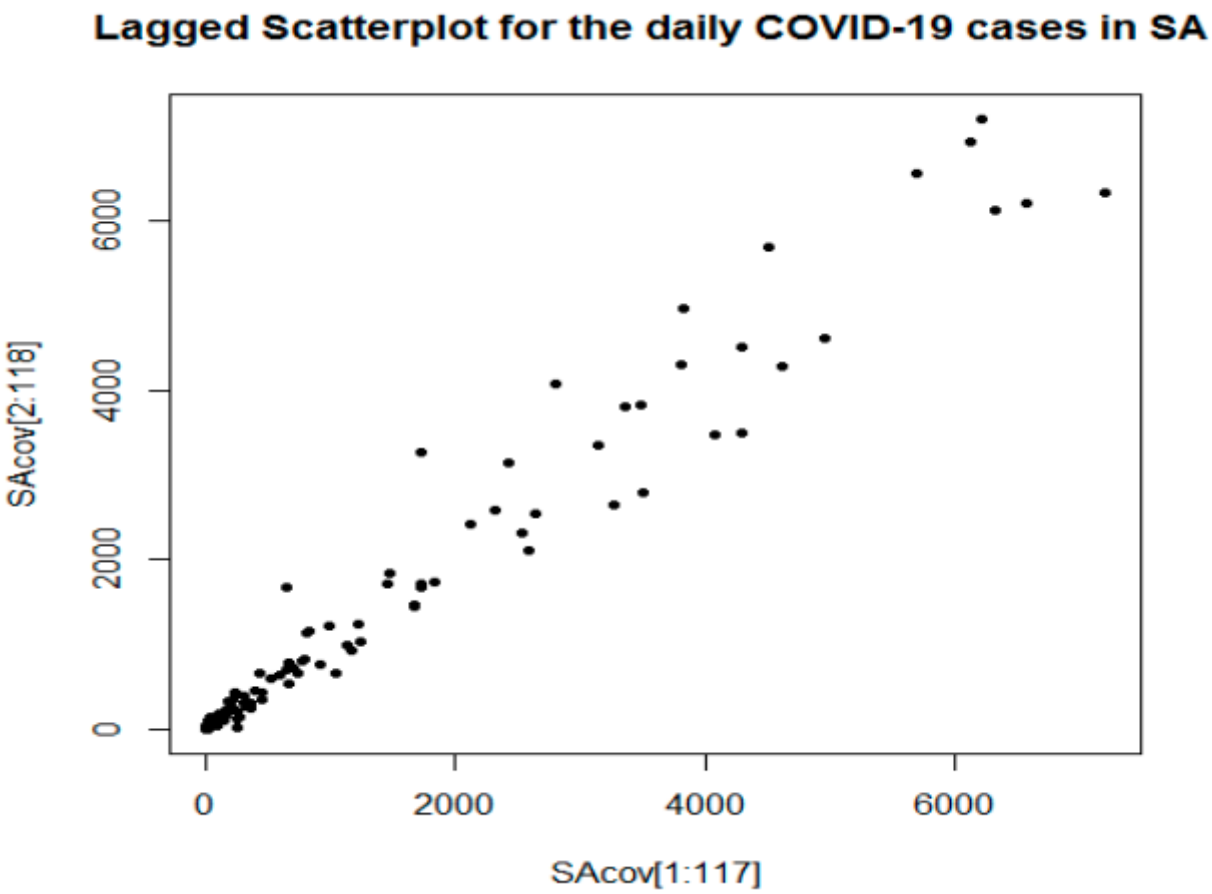

Figure 2

Lagged Scatter plot for the daily recorded cases for South Africa

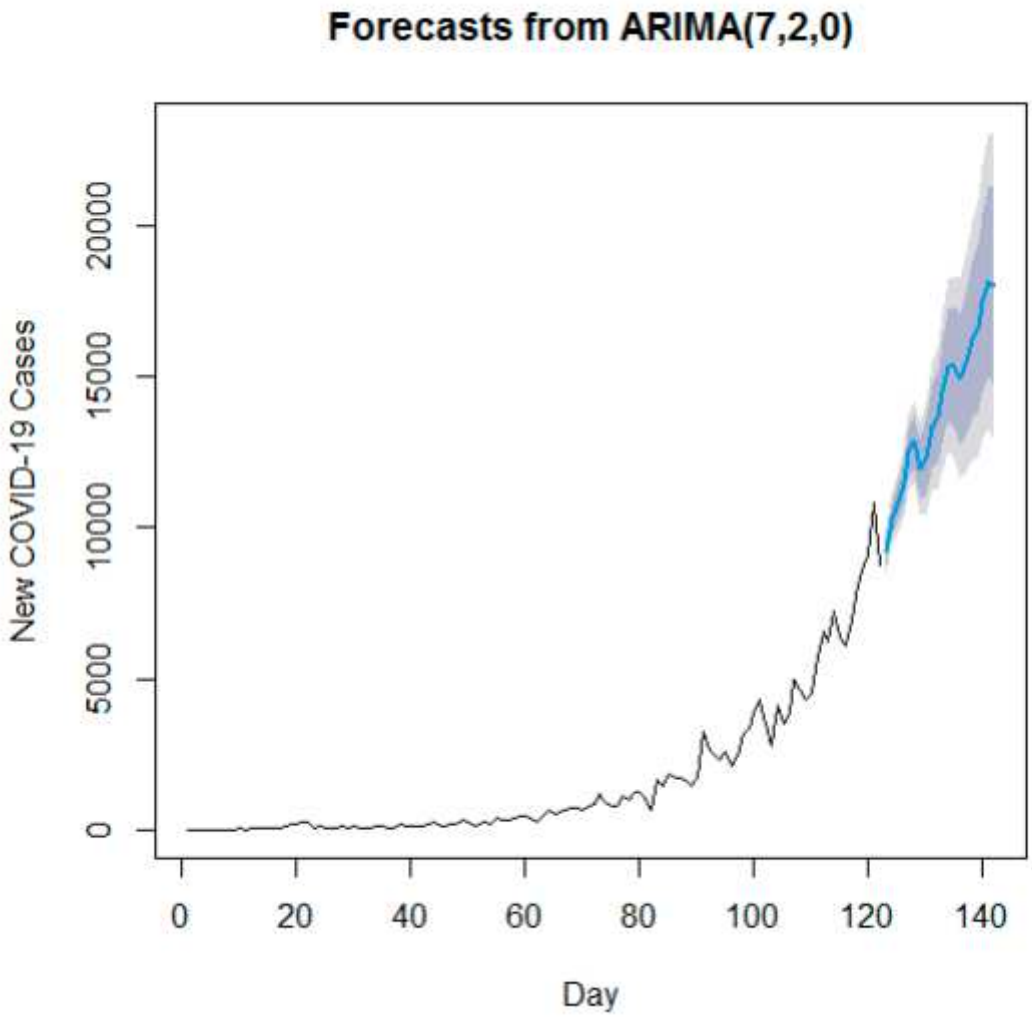

Figure 3

Forecasting plot of trajectory of daily COVID-19 cases in South Africa