International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2021.SP263

Scopus®

Original Research

IJCRR
Section: Healthcare
ISI Impact Factor (2019-20): 1.628
IC Value (2019): 90.81
SJIF (2020) = 7.893

Forecasting of Covid-19 Cases in India by Time Series Analysis Using Autoregressive Integrated Moving Average Model

Khobragade Ashish[1], Kadam Dilip[2]

[1]Assistant Professor, Department of Community & Family Medicine, All India Institute of Medical Sciences, Raipur, India; [2]Associate Professor, Department of Community Medicine, Seth G.S. Medical College, Mumbai, Maharashtra, India.

ABSTRACT

Introduction: COVID-19 is caused by SARS-CoV-2, a coronavirus. Forecasting has an important role in the surveillance of new emerging diseases like COVID-19.

Objective: The objective of the study was to forecast COVID-19 cases by using the ARIMA model.

Methods: We have used the ARIMA model to forecast cases of COVID-19 occurring per day in India. A total of 50 observations were used to fit the model. Model is best fitted by using order (0,2,1) which has the lowest AIC value. Forecasted values were compared with actual values.

Results: We have found that actual reported cases per day were within 95% CI of forecasted values.

Conclusions: ARIMA model can be used to forecast over a short period. This model can be used to develop strategies for the containment of pandemics.

Key Words: ARIMA, COVID-19, Forecast, India, Model, Time series analysis

INTRODUCTION

The first confirmed case of COVID-19 was reported from Wuhan city of China. Soon cases started spreading in China and nearby countries. It was announced as public health emergency of international concern (PHEIC) by the World Health Organization (WHO) in January 2020.[1] Alert was also given by the WHO regarding the global spread of the disease in March 2020.[1]More than 151.8 million cases and 31.86 lakh deaths were reported from all over the world as on 2nd May 2021. Total 19.8 million confirmed cases and 2.18 lakh deaths have occurred in India as of 2nd May 2021.[2]

If we can forecast the new cases which will occur in near future, it will aid in planning the resources required to prevent the occurrence of further cases. Many forecasting models were tried in the case of COVID-19 all over the world and in India.[3,4,5] In this study, we have tried a forecasting model based on autoregressive integrated moving average (ARIMA).

Objectives:

1. To develop a time series analysis (tsa) forecasting model of COVID-19 cases in India
2. To compare the actual cases of COVID 19 that occurred in India with the forecasted model.

Methodology:

The COVID-19 cases occurring daily in India were reported to the Government of India which has, in turn, appeared in the official dashboard. For the study purpose, the number of COVID-19 cases occurring daily from 17th February to 7th April 2021 per day in India was considered.[2,6] The data is available freely on the Government of India portal. This data was extracted in an excel sheet date wise. This data for 50 days was used to predict the occurrence of new cases of COVID-19 for the next five days. The epidemic curve was plotted for the selected period. [Figure 1]

In time series analysis, an ARIMA model is used to forecast future data based on the past available data. The Time series model is forecasted using the 'Forecast' package in R software. Fifty data points (daily COVID-19 cases) were

Corresponding Author:
Khobragade Ashish, Assistant Professor, Department of Community & Family Medicine, All India Institute of Medical Sciences, Raipur, India; Contact: 9834773698; Email: aw_k2008@rediffmail.com.

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 03.03.2021 **Revised:** 09.04.2021 **Accepted:** 06.05.2021 **Published:** 11.06.2021

first converted to time series data using the 'ts' command by taking the start point and endpoint of the data. The frequency was taken as 365.25 as data was related to daily COVID cases. Non-seasonal ARIMA models are generally denoted by ARIMA order (p,d,q). This order has three components to forecast the model: p, d and q, where p= number of autoregressive terms, d=order of differencing and q=number of lagged forecast errors in prediction.

The pre-requisite to apply the ARIMA model on time series data is that time-series data should be stationary. Time series data is called stationary when it's mean; variance and autocorrelation are constant over some time. Hence, the Augmented Dickey-Fuller (ADF) test was applied on time series data to check for its stationarity. Differencing was done twice to make time-series stationary. Consecutive numbers were subtracted twice for second-order differencing. In this way by differencing, we have removed the trend and seasonality of the data and hence, the mean of the time series is now constant. The stationarity of the data was confirmed by conducting the ADF test again. If the p value is less than 0.05, the data is considered stationary.

The order of the ARIMA model was selected by plotting autocorrelation (ACF) and partial autocorrelation (PACF) graphs. P-value was obtained from pacf graph and q value from acf graph.[Figure 2 and 3] ARIMA model was fitted by using the auto Arima function. Akaike's information criterion (AIC) was used for fitting the best model. Order having the lowest AIC value was selected as the best fitting model. Forecasting was done by using the fitted model for the next 5 days. The actual number of COVID-19 cases that occurred during this period was compared with forecasted data. The predicted model for the next 30 days was plotted graphically.

Statistical analysis: Statistical analysis was done by using R software version 3.6.1 using the 'Forecast' package.

Results: Original data was tested for its stationarity using Augmented Dickey-Fuller (ADF) test and the test results is as follows

ADF= 0.92, Lag order = 3, p-value = 0.99.

The data is not stationary as the ADF test p-value is more than 0.05. Hence, we have differenced the time series twice to make it stationary. After making differencing, ADF test results are as follows

ADF= -4.78, lag order=3, p value= 0.01. (p value < 0.05 is considered as significant)

The best fitted ARIMA model is (0,2,1) with the lowest AIC value of 965.79. The moving average (ma) coefficient for the fitted model is -0.8796 with a standard error of 0.0649. [Table 1].

The output of the forecasted model for the next 5 days is shown. The actual and predicted cases from 8th to 12th April 2021 is shown in table No.1. All the actual cases are within the range of 95% confidence interval of the predicted cases. [Table 2] Also predicted cases for the next 30 days (from 8th April to 7th May 2021) are plotted graphically [Figure 4].

DISCUSSION

In this study, we have used the ARIMA model to forecast cases of COVID-19. Without forecasting, it is very difficult to plan the strategies for the surveillance of the disease. When we have a forecasted data, public health surveillance can be carried out in the right direction and inculcate correct intervention measures. Hence, we have planned to do a time series analysis of COVID-19 cases to prove the hypothesis of whether these COVID-19 cases follows time series or not and to forecast the future trends.

We have taken COVID-19 cases that occurred in India as time series data of 50 days. As the data was not stationary, we have done differencing twice to take it stationary. We have used the ARIMA model to forecast using R software. The fitted model for that time series data in order (0,2,1). The AIC and BIC value is lowest for this order. ADF test was used to check stationarity. We have forecasted data for the next 5 days from 8th April to 12th April 2021. We have found that all of the actual cases reported are within the 95% confidence interval of the forecasted cases.

Different ARIMA models are fitted for different countries. In Saudi Arabia, the preferable ARIMA model is (2,1,1).[7] Best fitted model for various countries are Italy (0,2,1), Spain (1,2,0) and France (0,2,1).[8] Kabir et al. developed the ARIMA model for Nigeria. They used 39 observations to predict the corona cases.[9] We have used 50 days of data to predict the daily cases of COVID-19. Amal et al. forecasted the cases for 10 days using ARIMA and NARANN model. In this study, they used only 1-month of data to predict the cases.[10]A model to predict cases and deaths were developed in Italy to predict. In this study, the model was fitted by using order (0,2,0) and (2,2,1).[11] Similarly, one study forecasted COVID-19 cases for two days using ARIMA model of order (1,2,0) and (1,0,4).[12] In our study, ARIMA model is best fitted by using order (0,2,1). Actual cases reported for the next 5 days are within a 95% confidence interval of the predicted values.

Previously many time series models were used for forecasting infectious disease surveillance. Public health experts can predict how much variability will be there in future regarding the pattern of the disease.[13] Cases should be updated regularly so that if there is any change in the time trend of the disease, it will reflect in the model. The model will give a good prediction of the future trend of the disease. ARIMA model can be used for epidemiological surveillance of the new emerging diseases like COVID-19. So that correct intervention can be done at the correct time to prevent morbidity and mortality from the disease.

CONCLUSION

Actual cases of COVID-19 are within 95% CI of the predicted ARIMA model (0,2,1). ARIMA model can accurately predict the occurrence of COVID-19 cases. This model may be developed for state & district levels to predict COVID-19 cases.

Recommendation: Forecasting must be a part of routine surveillance activities in the pandemic situation of new emerging diseases like COVID-19.

ACKNOWLEDGEMENT

We express gratitude to the Ministry of Health & Family Welfare, Government of India for the supporting data, used in this research.

Conflict of interest: None

Source of funding: Nil

Author's contributions: Both authors have conceptualized the article. The manuscript was written by the 1st author and edited by the 2nd author. Data analysis was done by the 1st author.

REFERENCES

1. WHO. Timeline: WHO COVID-19 response. Available from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline
2. WHO. Coronavirus (COVID-19) Dashboard. Available from https://covid19.who.int/
3. Aravind M, Srinath K, Maheswari N, Sivagami M. Predicting COVID-19 Cases in the Indian States using Random Forest Regression. Int J Cur Res Rev. 2021;3:109-114.
4. Theerthagiri P, Jacob JI, Ruby UA, Yendapalli V. Prediction of COVID-19 Possibilities using K-Nearest Neighbour. Class Int J Curr Res Rev. 2017;3:156-164.
5. Yash S, Shikhar B, Parvathi R. Covid-19 Forecasting and Analysis Using Different Time-Series Model and Algorithms. Int J Cur Res Rev.2021;23: 184-189.
6. MOHFW, Govt. of India. COVID-19 status. Available from https://www.mohfw.gov.in/
7. Alzahrani SI, Aljamaan IA, Al-Fakih EA. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using the ARIMA prediction model under current public health interventions. J Infect Public Health [Internet]. 2020;13(7):914–9. Available from: https://doi.org/10.1016/j.jiph.2020.06.001
8. Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ. 2020:10;(7):138-148. Available from: https://doi:10.1016/j.scitotenv.2020.138817
9. Abdulmajeed K, Adeleke M, Popoola L. Online Forecasting of Covid-19 Cases in Nigeria Using Limited Data. Data Br. 2020; 30:105683. Available from: https://doi.org/10.1016/j.dib.2020.105683
10. Saba AI, Elsheikh AH. Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. Process Saf Environ Prot. 2020 Sep;141:1-8. doi: 10.1016/j.psep.2020.05.029.
11. Yang Q, Wang J, Ma H, Wang X. Research on COVID-19 based on ARIMA model$^{\Delta}$-Taking Hubei, China as an example to see the epidemic in Italy. J Infect Public Health. 2020 Jun 20:S1876-0341. doi: 10.1016/j.jiph.2020.06.019.
12. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Br. 2020; 29:105340. Available from: https://doi.org/10.1016/j.dib.2020.105340
13. Allard R. Use of time-series analysis in infectious disease surveillance. Bull World Health Organ. 1998;76(4):327–33.

Table 1: ARIMA Model for COVID-19 cases in India (0,2,1)

	ma1
Coefficient	-0.8796
Standard error	0.0649
sigma^2 estimated as 29195095: log likelihood= -480.9	

AIC=965.79 AICc=966.06 BIC=969.53

ma: moving average, AIC: Akaike's information criterion, AICc: Akaike's information criterion corrected, BIC: Bayesian's information criteria

Table 2: Forecast of COVID-19 cases using ARIMA model (0,2,1) and actual reported cases

Date	Point forecast	95% CI	Actual
08-04-2021	132771.4	122181.2 - 143361.6	131968
09-04-2021	138753.8	122850.2 - 154657.5	145384
10-04-2021	144736.2	124106.8 - 165365.6	152879
11-04-2021	150718.6	125552.4 - 175884.9	168912
12-04-2021	156701.1	127045.1 - 186357.0	161736

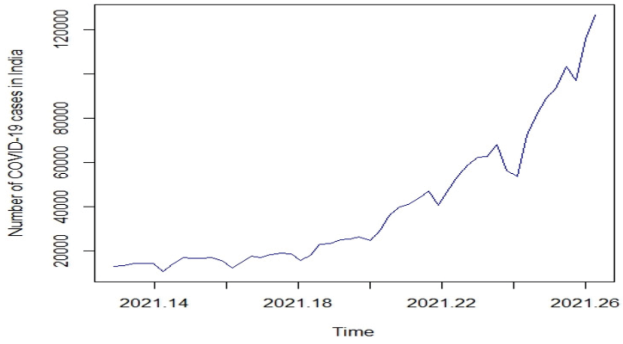

Figure 1: COVID-19 confirmed cases in India from 17th February to 7th April 2021.

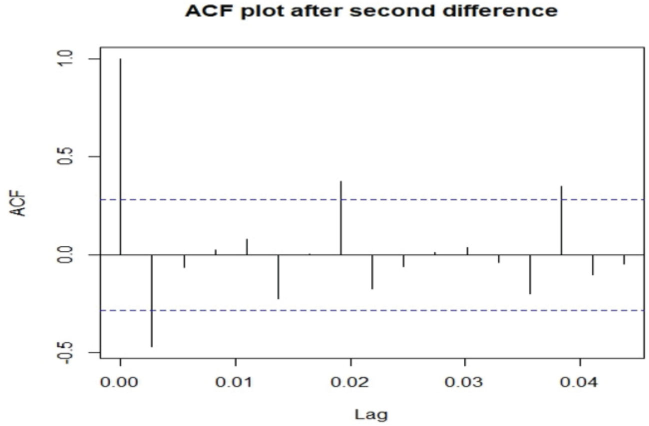

Figure 2: Auto-correlation graph after second order differencing.

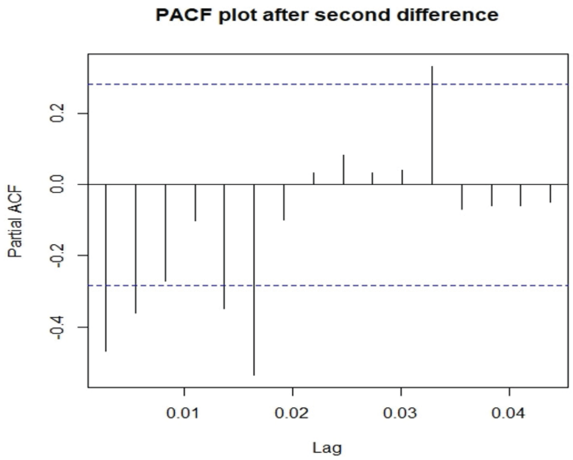

Figure 3: Partial auto-correlation graph after second order differencing.

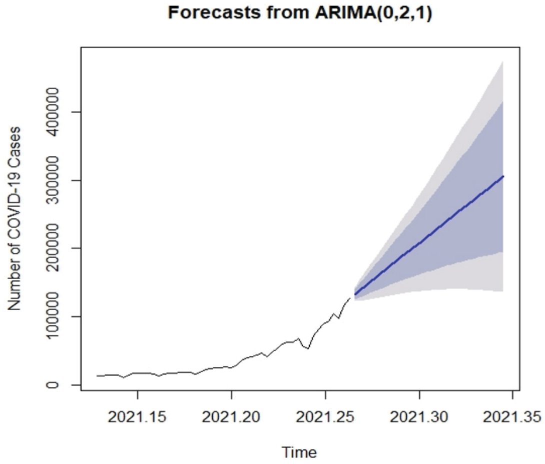

Figure 4: Forecast of COVID-19 cases for the next 30 days (From 8th April 2021 onwards).