REVIEW ARTICLE

Autoregressive Integrated Moving Average Model for Forecasting COVID-19 in India

Ariarathinam Newtonraj[1], Manikandan Mani[2]

Abstract

Introduction: We aimed to predict the trend of coronavirus disease-2019 (COVID-19) in India using autoregressive integrated moving average (ARIMA) model. Even though many models are available, ARIMA is found to be the best in COVID-19 situation; because ARIMA uses actual data and predicts accurately for a short period and thus proved as an effective public health tool especially in COVID-19 pandemic. On the contrary, most of the other models uses assumptions instead of actual data; based on agent, host, environmental factors, which is not yet clearly understood in this pandemic and those models will suit for well-understood diseases.

Materials and methods: The trend of COVID-19 was predicted using Gretl software and ARIMA model. Information about the COVID-19 was collected on a daily basis during January 1, 2020 to July 4, 2020 and predicted from July 5 to August 30, 2020. Various ARIMA models were assessed and the best one was selected. Then, the model's fitness was evaluated based on the normality of the residuals' distribution, based on the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). We also compared the first 20 days' data with actual data.

Results: The study results indicated that the ARIMA model (1,2,1) in confirmed cases could appropriately predict the cases and death in India and it is predicted that the cases may reach to 2.4 million by September 5, 2020 from the current status of 0.64 million.

Conclusion: Autoregressive integrated moving average prediction is a good tool in public health, for the current COVID-19 pandemic.

Keywords: Autoregressive integrated moving average, Corona, COVID-19, Forecast, Forecasting, Modeling.

Journal of Postgraduate Medicine, Education and Research (2020): 10.5005/jp-journals-10028-1395

Introduction

The novel coronavirus disease-2019 (COVID-2019) is a severe ongoing pandemic that has emerged in Hubei, a central province of China, in December 2019.[1] In few months, it has spread rapidly across the world, and it has been affected all the countries and has caused about 11 million confirmed cases, 0.5 million deaths as of July 4, 2020.[2] The severely affected countries are United States of America, Brazil, Russia, India, and United Kingdom.[2] COVID-19 is a pandemic caused by a newly discovered coronavirus and maximum number of people from all of the world infected with the new coronavirus (2019-nCoV) at the present time.[3–5] It can be easily transmitted humans to humans and the common symptoms of the disease are fever, fatigue, and dry cough.

In India, over the past few weeks, there have been several cases were reported and many people with exposure to suspected cases of COVID-19 and infected persons have also tried to dodge the mandatory home quarantine. In this condition, when the illness does not have any specific treatment, the prevention of disease and preparation in healthcare services are very important.[6] Statistical models are widely used to predict the spreading of the disease and figure out the number of cases and deaths,[7] which could help the health providers for strengthening resources, logistics enhancement for health workers to providing better service to the community and it can be useful for controlling this epidemic threat.[7–10] In this study, we adopted an autoregressive integrated moving average (ARIMA) model to forecast the number of new cases of COVID-19 for the upcoming months.

Materials and Methods

In this study, the COVID-19 data have been collected from the covid19india.org website, with the number of confirmed cases from January 1, 2020 to July 4, 2020.[11] Data have been divided into a training set (80%) on which our models are trained and testing set (20%) to test the performance of the model. Time series ARIMA model have been used to predict the future number of cases and the model consists of two components, the autoregressive (AR) model and moving average (MA) model.

The ARIMA model is usually denoted as ARIMA (p,d,q), where p is the order of AR component and q is the MA component and the d is the difference processing to the stationarity time series. The general equation of the ARIMA model is

$$y_t = \mu + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$$

where y_t is the value at time t, ϕ is the parameter of AR, and θ is parameter of MA. The stationary of the time series data is checked with augmented Dickey–Fuller (ADF) unit root test and autocorrelation function (ACF) and partial autocorrelation function (PACF). The better model was can be selected based on the minimal Akaike Information Criterion and mean square error of the selected

[1,2]Department of Community Medicine, Pondicherry Institute of Medical Sciences, Puducherry, India

Corresponding Author: Ariarathinam Newtonraj, Department of Community Medicine, Pondicherry Institute of Medical Sciences, Puducherry, India, Phone: +91 9489455994, e-mail: newton2203@gmail.com

How to cite this article: Newtonraj A, Mani M. Autoregressive Integrated Moving Average Model for Forecasting COVID-19 in India. J Postgrad Med Edu Res 2020;54(3):122–125.

Source of support:

Conflict of interest:

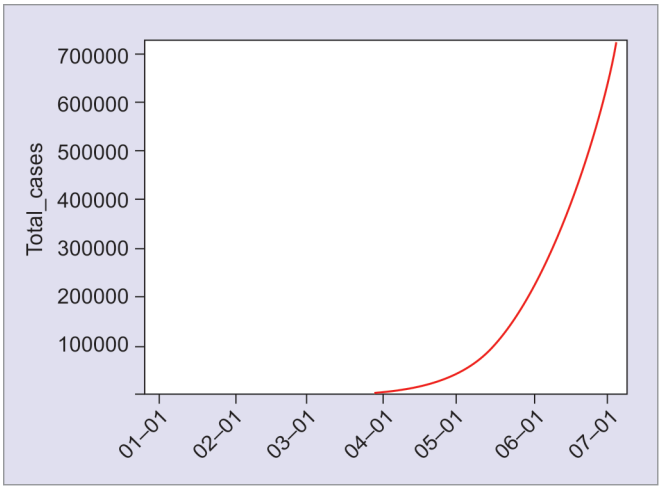

Fig. 1: Time series plot for confirmed cases of COVID-19 in India from January 1, 2020 to July 4, 2020

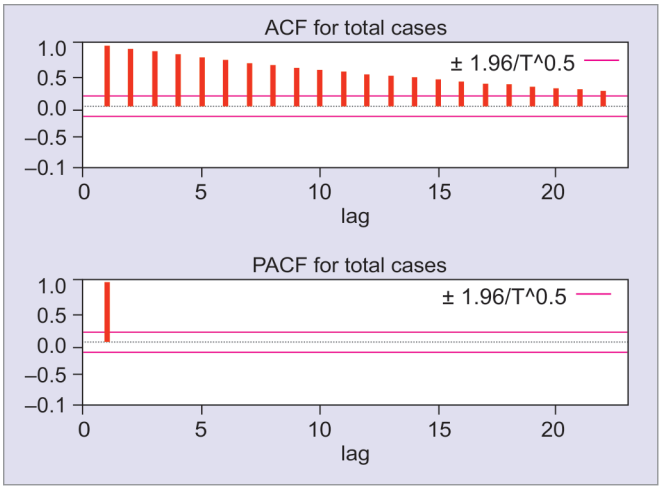

Fig. 2: Autocorrelation function (ACF) and partial autocorrelation function (PACF)

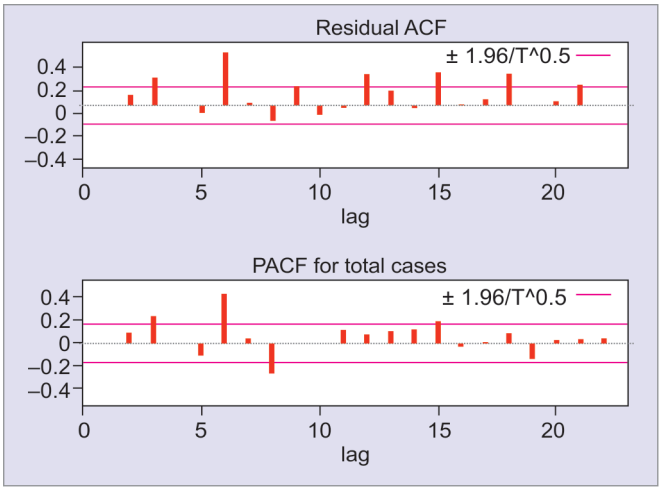

Fig. 3: Residual autocorrelation function (ACF) and partial autocorrelation function (PACF)

model. The estimate of the model would be diagnosed with residual ACF and PACF. Finally, the forecasting would be performed with the selected model. Institute administrative approval was obtained (dated July 7, 2020).

Results

The actual data show as of July 4, 2020, India reported around 0.64 million of confirmed cases, 0.2 million of active cases, 0.4 million of recovery cases, and 18,000 deaths.[11] Time series trend was identified using time series plot (Fig. 1), it shows the exponentially increasing trend in COVID-19 cases and our predictions shown the cases will reach 2.4 million by September 5. The stationarity was checked with ACF and PACF (Fig. 2) and the ADF test. It shows that the data are not stationary and ACF shows that exponentially decay and need to be transformed into a stationary series using order difference. The best ARIMA model was selected based on the minimum Akaike Information Criterion, MSE (Table 1). After selecting the better model, the parameters were estimated. Then, the stationarity of the residuals checked with residual ACF and PACF (Fig. 3). Finally, the selected ARIMA (1,2,1) was used to forecasting the number of cases for the next 3 months from July 5, 2020. The Gretl software was used to perform the analysis. Forecasting is shown in the Table 2 and the forecasting accuracy was checked by plot the forecasted data (Fig. 4).

Table 1: Model selection

Model	*AIC*	*R-squared*
ARIMA (1,2,1)	2010.03	0.99
ARIMA (1,1,1)	2045.68	0.93
ARIMA (2,1,1)	2040.39	0.95
ARIMA (2,1,2)	2030.36	0.96

Discussion

We used COVID-19 data to predict the number of cases in the future. The estimated model given the very accurate forecasting results. The time series model applied for the number of confirmed cases of COVID-19 in India and the time series trend was identified using the time series plot. The ARIMA model has been selected based on ACF, PACF, and the suitable information criterion and testing the assumption about the residuals of the estimating parameters.

Our prediction has the following important public health implications. First, our model predicts for next 2 months and we did not intent to predict further because COVID-19 is a novel virus and its transmission attributes are still under research. Predicting long-term transmission with its limited understanding on the transmission attributes would not deliver a meaningful results. Moreover, lockdown measures and preventive measures taken by the government are the main attributes in breaking the transmission cycle which is not a predefined one. So, a short-term prediction like ours will help the government and policy makers to prepare in line with their public health policy. Second, there are many prediction methods available and ARIMA is the most widely use method in medical statistics for prediction. This makes our prediction a more practical tool for public health use. Third, as ours is a short-term prediction, this will also takes care of the transmission-modifying attributes like lockdown measures and other preventive measures, mainly advocating of wearing masks, hand washing, maintaining

Table 2: Forecasting number of confirmed cases of COVID-19 in India with 95% confidence intervals from July 5, 2020 to September 30, 2020

Date	*Forecasted cases*	*CI*	*Date*	*Forecasted cases*	*CI*
06-07-2020	721,651	(720,818–722,483)	06-08-2020	1,535,729	(1,420,031–1,651,427)
07-07-2020	745,979	(743,945–748,012)	07-08-2020	1,564,067	(1,442,930–1,685,203)
08-07-2020	770,425	(766,856–773,994)	08-08-2020	1,592,535	(1,465,877–1,719,192)
09-07-2020	794,998	(789,618–800,378)	09-08-2020	1,621,132	(1,488,871–1,753,392)
10-07-2020	819,700	(812,271–827,128)	10-08-2020	1,649,859	(1,511,916–1,787,803)
11-07-2020	844,531	(834,842–854,220)	11-08-2020	1,678,717	(1,535,011–1,822,422)
12-07-2020	869,492	(857,349–881,635)	12-08-2020	1,707,704	(1,558,158–1,857,249)
13-07-2020	894,584	(879,808–909,359)	13-08-2020	1,736,820	(1,581,358–1,892,283)
14-07-2020	919,805	(902,229–937,380)	14-08-2020	1,766,067	(1,604,611–1,927,523)
15-07-2020	945,155	(924,623–965,688)	15-08-2020	1,795,444	(1,627,919–1,962,968)
16-07-2020	970,636	(946,996–994,276)	16-08-2020	1,824,950	(1,651,284–1,998,616)
17-07-2020	996,247	(969,358–1,023,136)	17-08-2020	1,854,587	(1,674,704–2,034,469)
18-07-2020	1,021,987	(991,712–1,052,262)	18-08-2020	1,884,353	(1,698,182–2,070,523)
19-07-2020	1,047,858	(1,014,065–1,081,650)	19-08-2020	1,914,249	(1,721,719–2,106,779)
20-07-2020	1,073,858	(1,036,421–1,111,294)	20-08-2020	1,944,275	(1,745,314–2,143,235)
21-07-2020	1,099,988	(1,058,785–1,141,190)	21-08-2020	1,974,431	(1,768,969–2,179,891)
22-07-2020	1,126,248	(1,081,161–1,171,334)	22-08-2020	2,004,716	(1,792,685–2,216,747)
23-07-2020	1,152,638	(1,103,551–1,201,723)	23-08-2020	2,035,132	(1,816,462–2,253,801)
24-07-2020	1,179,157	(1,125,960–1,232,354)	24-08-2020	2,065,677	(1,840,301–2,291,052)
25-07-2020	1,205,807	(1,148,390–1,263,223)	25-08-2020	2,096,352	(1,864,203–2,328,501)
26-07-2020	1,232,586	(1,170,843–1,294,328)	26-08-2020	2,127,157	(1,888,168–2,366,146)
27-07-2020	1,259,495	(1,193,323–1,325,666)	27-08-2020	2,158,092	(1,912,197–2,403,987)
28-07-2020	1,286,534	(1,215,832–1,357,235)	28-08-2020	2,189,157	(1,936,291–2,442,023)
29-07-2020	1,313,703	(1,238,372–1,389,033)	29-08-2020	2,220,352	(1,960,450–2,480,253)
30-07-2020	1,341,002	(1,260,945–1,421,058)	30-08-2020	2,251,676	(1,984,675–2,518,678)
31-07-2020	1,368,430	(1,283,553–1,453,307)	31-08-2020	2,283,131	(2,008,966–2,557,295)
01-08-2020	1,395,989	(1,306,198–1,485,780)	01-09-2020	2,314,715	(2,033,324–2,596,106)
02-08-2020	1,423,677	(1,328,881–1,518,473)	02-09-2020	2,346,429	(2,057,749–2,635,109)
03-08-2020	1,451,496	(1,351,605–1,551,386)	03-09-2020	2,378,273	(2,082,243–2,674,303)
04-08-2020	1,479,444	(1,374,370–1,584,517)	04-09-2020	2,410,247	(2,106,805–2,713,689)
05-08-2020	1,507,522	(1,397,178–1,617,865)	05-09-2020	2,442,351	(2,131,436–2,753,265)

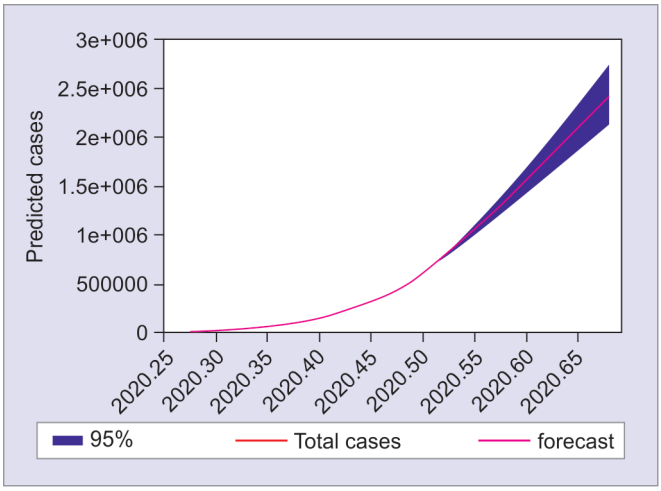

Fig. 4: Forecasting number of confirmed cases of COVID-19 with 95% confidence intervals from July 5, 2020 to September 5, 2020

social distance, and personal hygiene by the government. Fourth, prediction based on SEIR models uses more assumption which may lead to the failure of COVID-19 prediction, whereas ARIMA uses only the actual data and no assumptions and hence it has great public health value in a novel viral pandemic like COVID-19.[12–14]

In India, initial rise in cases was very slow due to strict lockdown and other factors and the lockdown has been slowly relaxed with restrictions.[8,11,15] This relaxation was also varied from state to state and depending on the case burden. In India, Mumbai, Delhi, and Chennai were the main metro cities first affected reporting large number of cases and the spread to other cities, smaller towns, and villages is low and at present this is slowly spreading.[11,15] Encouragingly, the mortality rate in India is 2.87% and is very low than many other countries in the world.[11] Appropriate statistical modeling like ours could play a key role in almost all aspects, including infection detection, traffic management, logistics supply chain, etc. In the current situation, number of new cases and death are increasing, which clearly shows that the outbreak of COVID-19

may devastating and it could result more infections and deaths across the country. Though, it could be controllable if the effective control measures are followed, several essential tasks must be continued to control the virus sources by keeping any suspected infected people to quarantine home or hospital and preventive measures should be continued in the general population.

Conclusion

The main advantages of ARIMA forecasting could help in the decision-making process based on the historical data. The results show that number of new cases would be increasing in the coming months and this suggests that if there was no major implementation on control measures like social distancing, lockdowns, stay-home-advises/orders, wearing masks in public place/workplace, and frequently use of hand sanitizer/hand wash then the virus will spread exponentially.

References

1. Zhang X, Ma R, Wang L. Predicting turning point, duration and attack rate of COVID-19 outbreaks in major western countries. Chaos Solitons Fractals 2020;135:109829. DOI: 10.1016/j.chaos.2020.109829.
2. Shortbread Developers. Covid 19 Info [Internet]. [cited 2020 Jul 4]. Available from: https://covid19info.live/.
3. Tandon H, Ranjan P, Chakraborty T, et al. Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. 2020. 1–11. arXiv:2004.07859.
4. Kumar P, Singh RK, Nanda C, et al. Forecasting COVID-19 impact in India using pandemic waves nonlinear growth models. medRxiv 2020. DOI: 10.1101/2020.03.30.20047803.
5. Dehesh T, Mardani-Fard HA, Dehesh P. Forecasting of COVID-19 confirmed cases in different countries with ARIMA models. medRxiv 2020. DOI: 10.1101/2020.03.13.20035345.
6. Perone G. An ARIMA model to forecast the spread of COVID-2019 epidemic in Italy. SSRN Electron J 2020. DOI: 10.1101/2020.04.27.20081539.
7. Anne WR, Jeeva SC, World T, et al. ARIMA modelling of predicting COVID-19 infections. Non Stationari 2020. 5–8. DOI: 10.1101/2020.04.18.20070631.
8. Sujath R, Chatterjee JM, Hassanien AE. A machine learning forecasting model for COVID-19 pandemic in India. Stoch Environ Res Risk Assess 2020(7). DOI: 10.1007/s00477-020-01827-8.
9. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons and Fractals [Internet]. 2020;134:109761. DOI: 10.1016/j.chaos.2020.109761.
10. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PLoS One [Internet] 2020;15(3):1–8. DOI: 10.1371/journal.pone.0231236.
11. A Crowdsource Initiative. COVID 19 India [Internet]. [cited 2020 Jul 4]. Available from: https://www.covid19india.org/.
12. Iwata K, Miyakoshi C. A simulation on potential secondary spread of novel coronavirus in an exported country using a stochastic epidemic SEIR model. J Clin Med 2020;9(4):944. DOI: 10.3390/jcm9040944.
13. Hou C, Chen J, Zhou Y, et al. The effectiveness of quarantine of Wuhan city against the corona virus disease 2019 (COVID-19): a well-mixed SEIR model analysis. J Med Virol 2020;92(7):841–848. DOI: 10.1002/jmv.25827.
14. Li MY, Graef JR, Wang L, et al. Global dynamics of a SEIR model with varying total population size. Math Biosci 1999;160(2):191–213. DOI: 10.1016/S0025-5564(99)00030-9.
15. Lancet T. India under COVID-19 lockdown. Lancet (London, England) 2020;395(10233):1315. DOI: 10.1016/S0140-6736(20)30938-7.