Research Article

ARIMA model for COVID-19 and its prediction in India

Abstract

In this paper, autoregressive integrated moving average (ARIMA) model has been applied to know the trend and to predict future pattern of present COVID-19 in India. Accuracy of the model has been checked. Data from July 1, 2020 to August 3, 2021 has been taken for the study. We estimated ARIMA model to forecast the epidemic trend over the period after July 1, 2020, by using the Indian epidemiological data (www.COVID19india.org)[1] at national level. The data refer to the number of daily confirmed, total confirmed and deceased cases officially registered by the Indian Ministry of Health (https://api.COVID19india.org/) for the considered period. The main aim of this study is to know the trend of COVID-19 daily cases as well as deceased cases, and forecast for next 120days after August 3, using appropriate ARIMA model.

Keywords: ARIMA, COVID19, KPSS-test, PP-test, MSE, ME, Confirmed cases

Volume 10 Issue 4 - 2021

Kamlesh Kumar Shukla,[1,2] Syed Azizur Rahman,[3] Ranjana Singh,[4] Rama Shanker[5]

[1]Professor, Department of Mathematics, School of Sciences, Noida International University, G.B. Nagar, India
[2]Department of Community Medicine, Noida International Institute of Medical Sciences, Noida International University, G.B. Nagar, India
[3]Assistant Professor, Department of Medicine, Noida International Institute of Medical Sciences, Noida International University, G.B. Nagar, India
[4]Professor & HOD, Department of Community Medicine, Noida International Institute of Medical Sciences, Noida International University, G.B. Nagar, India
[5]Associate Professor, Department of Statistics, Assam University, Silchar, India

Correspondence: Kamlesh Kumar Shukla, Professor, Department of Mathematics, School of Sciences, Noida International University, G.B. Nagar, India,
Email kkshukla22@gmail.com

Received: October 18, 2021 | **Published:** December 15, 2021

Introduction

The pandemic COVID-19 is initially recognized in China and could be traced back to a cluster of severe pneumonia cases identified in Wuhan, China in December 2019,[2,3] and after two months it starts spreading in the world. Almost all countries in the world are affected by COVID-19 and still going on its transmission from person to person, city to city, nation to nation. It has been affecting not only the health of persons but social- economic status of person, city as well as nation.

As we all know that a COVID-19 case has been started in India approximately in the month of January 2020 and thereafter spreading rapidly. Spread of this disease through human to human contact, a large number of cases have been reported worldwide by Hamzah, et al.[4] Due to the outbreak of this pandemic, an obligatory situation was formed, whereby the authorities of various countries and continents had to put restrictions on the movement of people and non-essential activities. Some of these restrictions included imposing of lockdowns, maintaining social distancing, and work from home in academics and in business continuity plans. Thus, the spread of COVID-19 has left a major impact on the environment as well as on the lifestyle of human beings.[5,6] Almost all academic institutions were closed and people were advised to work from home, and perform contactless financial transactions using various digital platforms,[7,8] In this way socio-economic status of a country has been affected.

The total number of confirmed, active, recovered and deceased cases were 203,456,777, 16,375,542, 182,773,747 and 4,307,488 recorded in the world as well as 3,19,69,596, 3,96,694 3,11,31,926, and 4,28,339 were recorded in India till August 8, 2021. As far as India is concerned, it is very challenging task to control/manage COVID-19, where the population is around 1.38billion. As we know that India has produced two vaccinations and started vaccinate to the people from January, 2021 onward. Although there is some restrictions and limitations of age of persons due to shortage of vaccinations, but Vaccination may also have impact in reducing the cases of novel corona virus.

Many researchers have studied on COVID-19 cases and applied different forecasting models such as Trigonometric Exponential Smoothing State Space model with Box-Cox transformation (TBAT) developed by De Livera, et al.[9] uses a combination of Fourier terms with an exponential smoothing state space model and a Box-Cox transformation,, Multiple Regression, ARIMA etc., but we applied optimum ARIMA model to predict future cases for four months because it was observed from the literature that ARIMA model is good model in terms of projection and it is having less error in comparison to other. To understand the future spread of pandemic and to devise management strategies, various models have been designed, which give information regarding the time of attainment of infection peak, the number of infected cases and the requirement of medical infrastructure to manage the spread.[10]

The main objectives of this study are: It can be easily identified the trend of COVID19 in India using ARIMA models. Its trend and forecast will be benefited to all governmental institutions, especially in public health for making further policy and plan. This study may be used to monitor the epidemic and to better allocation of the resources. Further useful and more precise forecasting may be provided by updating this study.

A study about COVID-19 is carried out in the present study and it is organized into four sections with section one as introductory in nature. Data and methodology have been discussed in the second section. In the third section, selected appropriate ARIMA model have illustrated using different statistical tools as well as appropriate graphs. Conclusions have been drawn on the basis of used analysis and reported in the fourth section.

Data and methodology

Descriptions about data and methods have been discussed in this section.

Data

The data have been taken for the analysis are the number of new daily COVID-2019, total confirmed cases and new daily deceased cases from July 1, 2020 to August 3, 2021, and are extracted from the official website of the Indian Ministry of Health (https://api.COVID19india.org/). They include the overall national trend and all states of India. Some states of India have been reported more cases in comparison to other states. As is evident from the data, the state of Maharashtra is the worst affected in terms of total cases which accounts for about 23% of the cases in India. The next four worst affected states/ union territories are Tamil Nadu, Andhra Pradesh, Karnataka and Delhi having approximately 38% of the total cases and the rest of Indian states/ union territories having another 39% cases. Using appropriate ARIMA model, new daily confirmed cases, total confirmed cases and daily decease cases have been projected after August 3, 2021.

Methods

From the few months literature, it is observed that several studies have been conducted on COVID-19 and its projection in different places, of which some important contributors are Batista 2020[11-19] have been studied using different models and approaches. In this study, ARIMA model has been applied to know the pattern and as well as to forecast for future cases of COVID-19 (daily confirmed, total confirmed cases and daily deceases cases), as one of study has already reported that ARIMA is the good model to forecast time series data.[12,15,16] It could be considered one of the good prediction models for time series.

The appropriate ARIMA(p, d, q) model parameters have been selected by using following criteria (i) the Akaike's information criterion (AIC); (ii) autocorrelation function (ACF) and partial autocorrelation function (PACF) of the residuals; and (iii) testing the general statistical assumptions about residuals. Using above criteria, ARIMA (3,1,2), ARIMA (3,2,2) and ARIMA (0,1,2) have been applied to project daily confirmed cases, total confirmed cases and daily decease cases respectively.

In the first stage, Phillips–Perron Unit Root test (PP-test) for unit root as well as Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test have been applied to check the stationary time series data and then ACF and PACF plots of residuals as well as QQ-plots, histogram and other residuals plots are used to select the appropriate ARIMA model. Mean Error (ME), Mean absolute error (MSE), Mean Absolute Standard Error (MASE) and Autocorrelation function (ACF) have been calculated from forecast models to check the accuracy of the model and Durbin Watson test has also been applied to check autocorrelation disturbances among the residuals.

Analysis and discussions

In this section, common statistical techniques related to ARIMA models have been discussed and importance of results has also been reported graphically. Present pattern of daily confirmed, total confirmed cases and daily deceased cases of COVID-19 are presented in the Figure 1, 2 & 3, respectively.

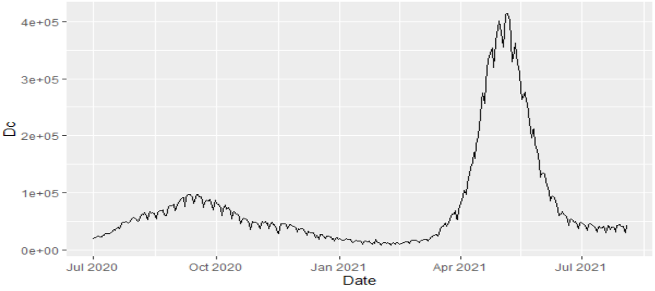

Figure 1 Daily confirmed cases (Dc) from July 1, 2020 to August 3, 2021).

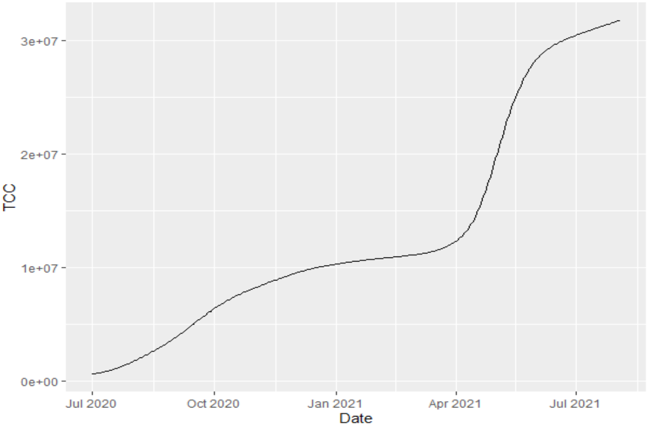

Figure 2 Total confirmed cases (TCC) from July1, 2020 to August 3, 2021.

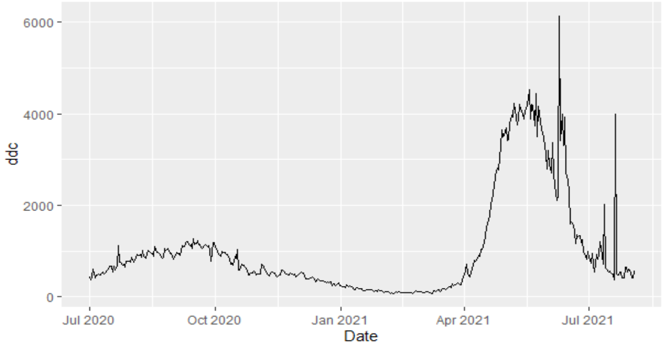

Figure 3 Daily deceased cases (dcc) (from July 1, 2020 to August 3, 2021).

From Figure 1, it is observed that daily confirmed cases were increased in the month of August & September, 2020 and it was reported almost 100,000 cases in a day, and again cases were increased as second waive in the month of May& June, 2021 and it was reported more than 400,000 cases in a day. Similar interpretation can be seen in the Figure 3 whereas Figure 3 shows about daily deceased cases from July 1, 2020 to August 3, 2021.

Figure 2 indicates about total confirmed cases (from July 1, 2020 to August 3, 2021) and it can been seen from the figure that total confirmed cases have been increasing continuously although there is rapid outbreak of COVID19 cases in the month of September 2020 (reported more than 10,000,000 cases) as well as fast rapid outbreak May 2021(reported more than 30,000,000cases), It is observed that it has increased more than 3times within 7months.

Phillips–perron unit root and kpss test

The stationarity of the time series data can be checked with the help of Phillips–Perron (PP) Unit Root test, given in Paron.[20] We used PP test to check about stationary and non-stationary time series and its results are given in Table 1. The p-value of the PP test is greater than the level of significance i.e., 0.05. Table 1 indicates that daily confirmed, total cases and daily deceased cases are non-stationary at 5% level of significance. It means all data (three cases) are required to transform into stationary time series data. Difference of series can be used to convert into stationary time series, where appropriate ARIMA models were selected with the help of minimum values of AIC, BIC and ME.

Table 1 Summary statistics and PP-test for COVID-19 cases

Variable	Mean	SD	Min.	Max.	PP- test	
					PP	p-value
Daily cases	78150	90315.87	8579	414280	-2.598	0.951
Total cases	13065966	9487445	605224	31767766	-0.581	0.99
Daily death	1022	1122.62	75	6139	-14.19	0.315

The KPSS test can be applied to know nature of time series, whether it is stationary or not. This test is proposed by Kwiatkowski et al.[21] The test p-value = 0.01, so we can assume that the series is not stationary. To check the stationarity of the series, we should have differences of the time series for the selection of appropriate model, difference of the series are required, It can be seen in the Figure 7, where 1st, 2nd and1st differences have been required to convert into stationary for daily confirmed, total confirmed and daily deceased cases respectively.

ARIMA models and estimation of parameters

Appropriate ARIMA models have been selected using Akaike information criteria (AIC), Baysian information criteria (BIC), Mean absolute percentage error (MAPE) and Mean Error (ME) values of the parameters of model, which are presented in the Table 1, 2 &3 respectively. Estimation of parameters of the selected models has been presented in the Table 4. Accuracy of models have also been verified using ACF and PACF plots of residuals for daily confirmed , total confirmed and total deceased cases , which are presented in the Figure 4,5 & 6 respectively. The absence of significance residuals spike indicates that almost all models are good fit. It can be seen from the Figure 14, 15 &16 for daily confirmed, total confirmed and daily decease cases respectively, where QQ-plots, histogram and scatter plots of residuals are presented. Except some variation in normal plots, rest plots have been indicated good fits for the selected ARIMA models.

Table 2 Summary of KPSS-test for COVID-19 cases

Cases	KPSS level	Lag parameter	p-value
Daily confirmed	1.1224	5	0.01
Total confirmed	5.8314	5	0.01
Daily deceased	1.5355	5	0.01

Table 3 The optimal ARIMA models for different variables

Cases	Parameters	AIC	BIC	MAPE	ME
Daily confirm	(3,1,2)	8245.01	8268.93	-0.32973	38.394
Total confirm	(3,2,2)	8225.06	8248.96	0.00711	33.2066
Daily Deceases	(0,1,2)	5774.28	5786.24	-3.62189	0.37571

Table 4 Estimation of parameters of ARIMA models for different variables

Cases	ARIMA	Parameters				
	Model					
		AR1(S.E.)	AR2(S.E.)	AR3(S.E.)	MA1(S.E.)	MA2(S.E.)
Daily confirm	(3,1,2)	0.0613(.05)	-0.7671(.02)	0.4068(.05)	0.1465(.03)	0.9397(.03)
Total confirm	(3,2,2)	0.0626(.04)	-0.7673(.03)	0.4062 (.04)	0.1438(.03)	0.9420(.02)
Daily Decease	(0,1,2)	-	-	-	-0.6199(.05)	0.0951(.05)

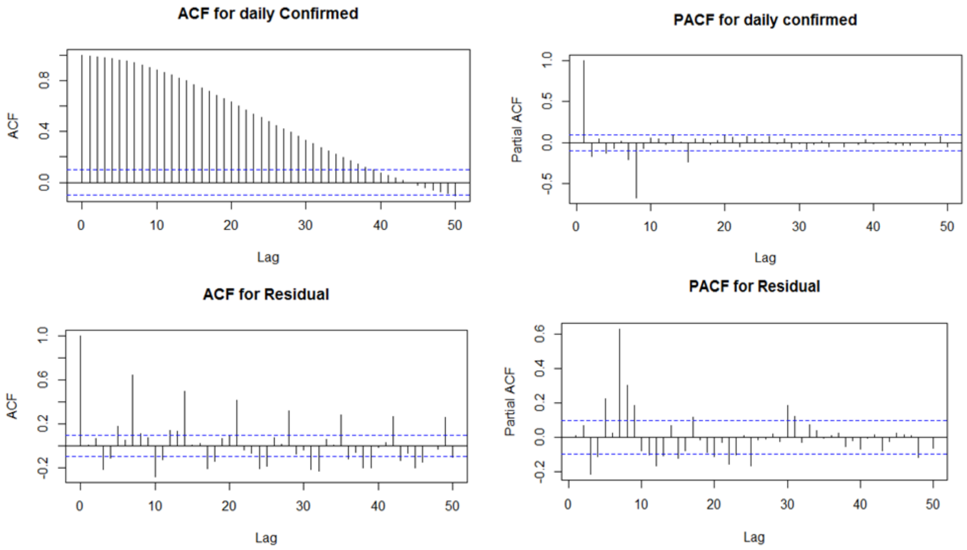

Figure 4 ACF and PACF plot of daily confirmed cases and its residuals.

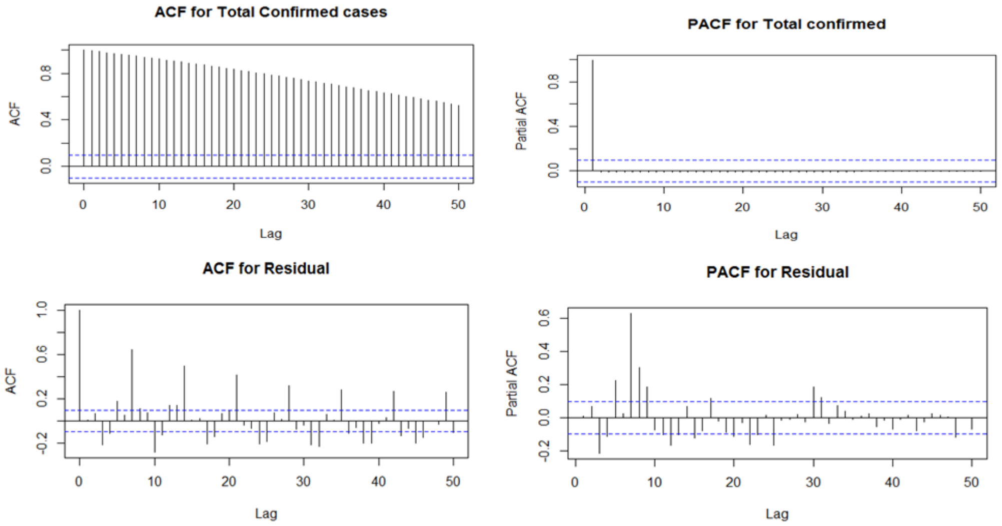

Figure 5 ACF and PACF plot of total confirmed cases and its residuals.

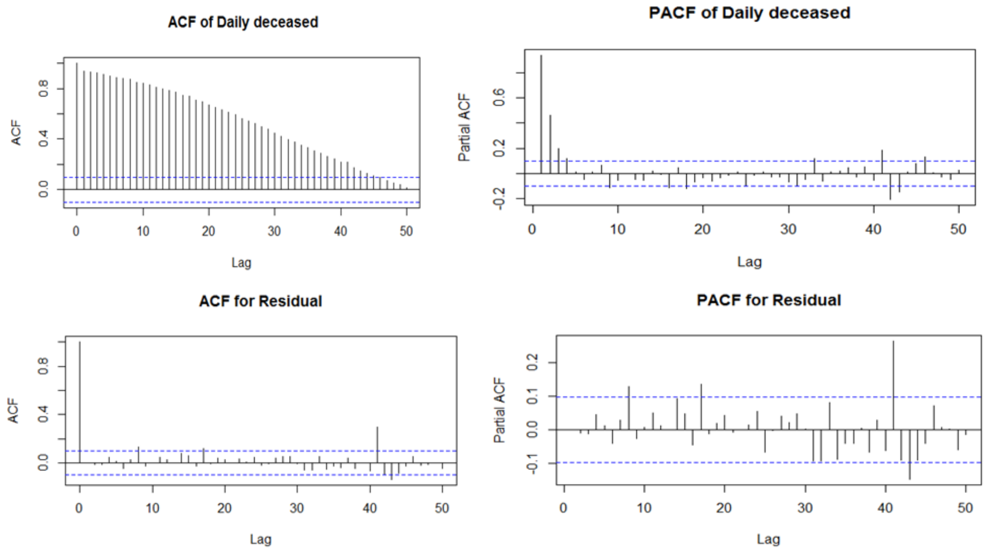

Figure 6 ACF and PACF plot of daily deceased cases and its residuals.

Fitted and observed values for daily confirmed, total confirmed and daily decease cases are presented in Figures 8,9 &10 respectively. From the figures, It is observed that very closed fit between observed and fitted values using ARIMA (3,1,2), ARIMA (3,2,2) and ARIMA (0,1,2) for daily confirmed, total confirmed and daily decease cases respectively.

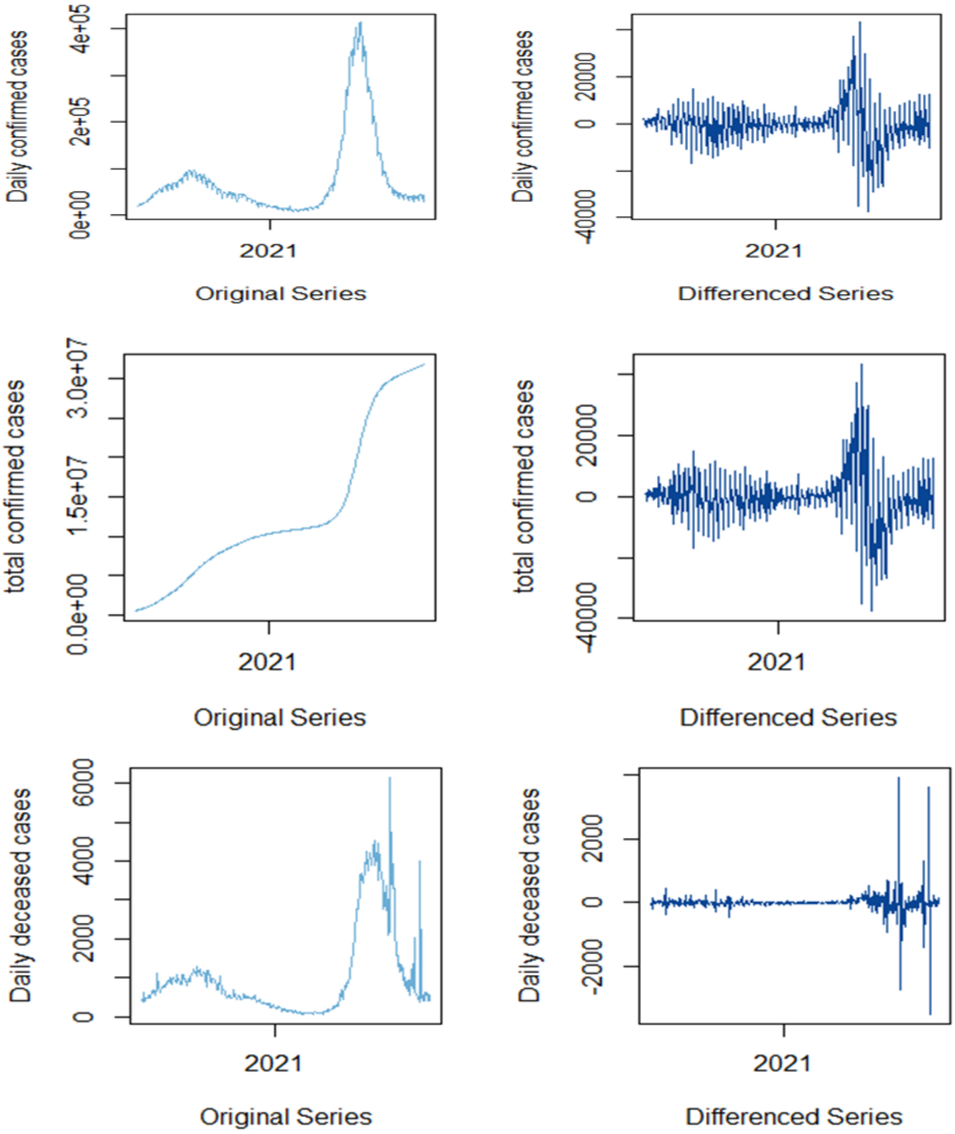

Figure 7 Plots of original and differences of daily confirmed, total confirmed, daily deceased cases respectively (from July, 2020, to August 3, 2021).

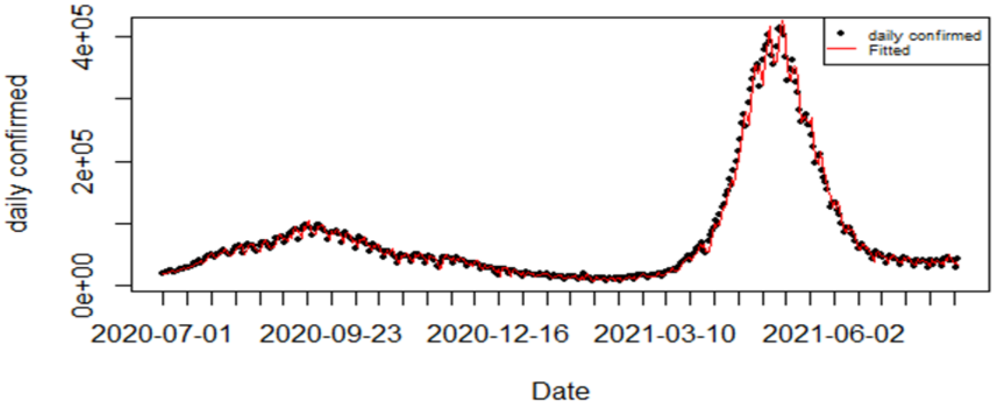

Figure 8 Fitted plot for daily confirmed cases using ARIMA (3,1,2).

Citation: Shukla KK, Rahman SA, Singh R, et al. ARIMA model for COVID-19 and its prediction in India. *Biom Biostat Int J.* 2021;10(4):176‒183. DOI: 10.15406/bbij.2021.10.00345

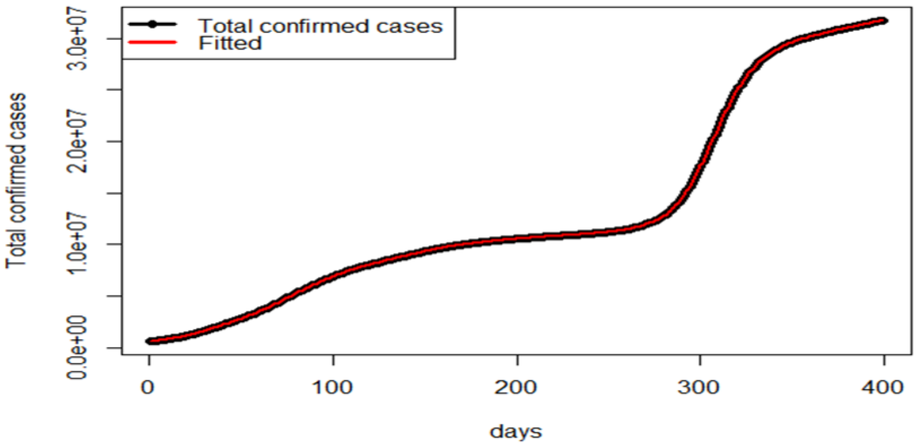

Figure 9 Fitted plot for total confirmed cases using ARIMA (3,2,2).

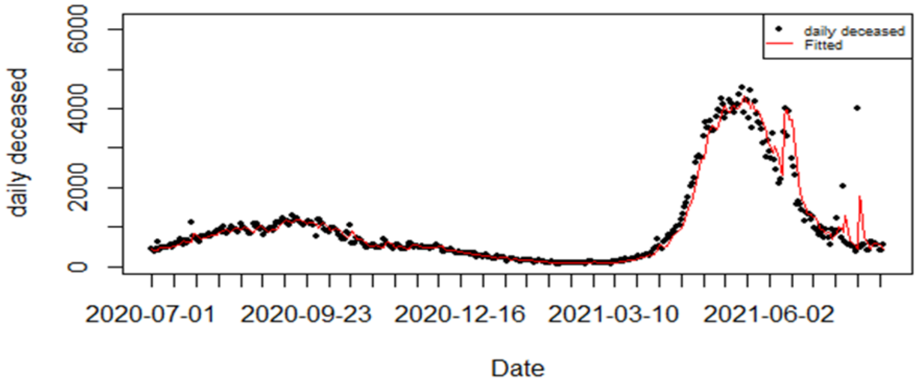

Figure 10 Fitted plot for daily deceased cases using ARIMA (0,1,2).

Figure 11 shows that forecast of ARIMA (3,1,2) for daily confirmed cases for next 120days (up to November 30, 2021), it is observed that daily cases are slightly decreasing with increased days. Point estimations and interval estimation (80% &95%) are calculated using optimum ARIMA model which indicates that there is no chances to increase daily confirm cases with increased days up to 512days which is up to November, 2021 as shown in the Figure 11. Details about the forecast (point & Interval estimation) along with the date wise for next 120days (from August, 4 to November 30, 2021) are mentioned in the Appendix table 1. Its interval estimations are increasing with increased days and it is due to past two waives of COVID-19, where daily cases were found very high during second waives. On the basis of first waives of COVID-19 as well as point estimation of the model, there may be no chances to increase daily cases with increased number of days, if situation remains the same. As we see that different states and territories of governments of India are announcing to open colleges, schools and malls from August/ September, 2021, there may be less chance to increase the daily confirm cases during November/ December, 2021.

Figure 12 indicates the forcasst for total confirm cases for next 120days (from August 4 to November 30, 2021) using ARIMA (3,2,2) model. Total confirmed cases are slightly increasing with increased days as shown in the figure.Total confirm cases have been found incrasing pattern because daily confirm cases are adding. As per pattern of Figure 12, total cases may be increased up to February/ March, 2022. If the daily confirm cases does not increase, it will become stationary.

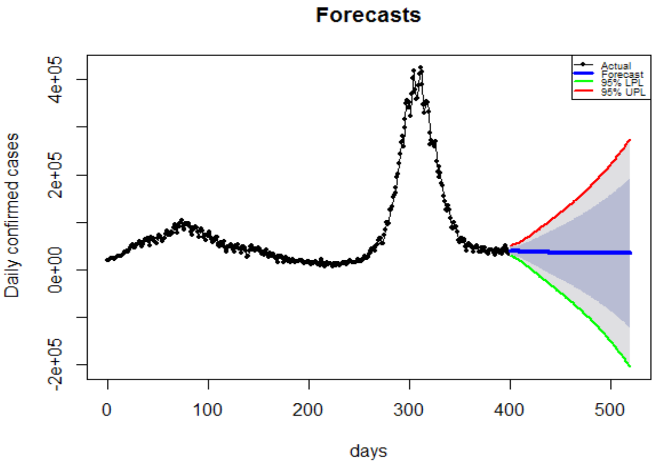

Figure 11 Forecast for next 120 days after 400 days for daily confirme cases.

Figure 13 shows the forecast for daily decease cases for next 120 days (August 4 to November 30, 2021) using ARIMA (0,1,2) model. Daily deceases case shows slightly decreasing pattern up to Novmber, 2021 according to the model.

Accuracy of models and its forecasts

Accuracy of models can be checked using the plots of residuals and its normality test. It can be seen from the Figure 14, 15 &16 respectively that the residual plots are normally distributed except

Citation: Shukla KK, Rahman SA, Singh R, et al. ARIMA model for COVID-19 and its prediction in India. *Biom Biostat Int J.* 2021;10(4):176–183. DOI: 10.15406/bbij.2021.10.00345 with a little variation in QQ-plot. Durbin Watson test has also been applied to check its normality and autocorrelation of residuals for all three cases as well as accuracy of forecasting of models, which are presented in the Table 5. If the value of autocorrelation is 2.00, it means there is no autocorrelation between residuals, if its value is less than 2 then positive autocorrelation and if it is more than 2 then negative autocorrelation between residuals using Durbin Watson. Mean Error (ME), Mean absolute standard error (MASE) and Auto correlation function (ACF) of error have been presented in Table 5.

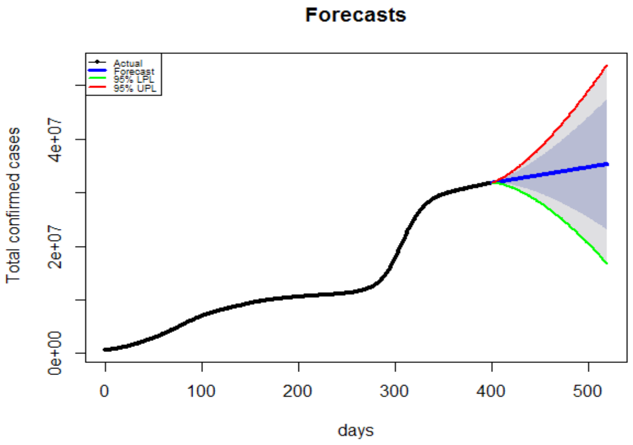

Figure 12 Forecast for next 120 days after 400days for total confirmed cases.

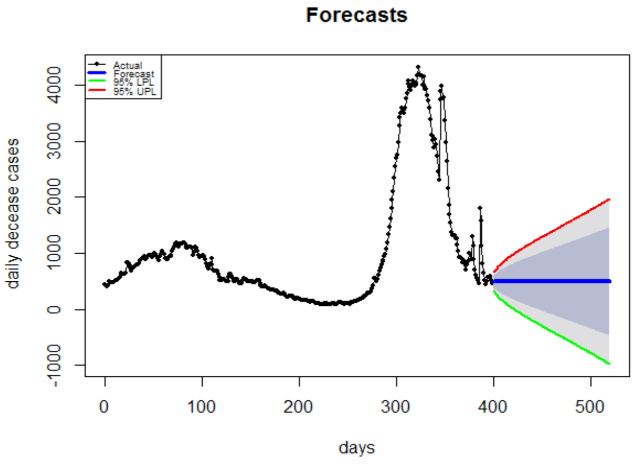

Figure 13 Forecast for next 120 days after 400days for daily decseas cases.

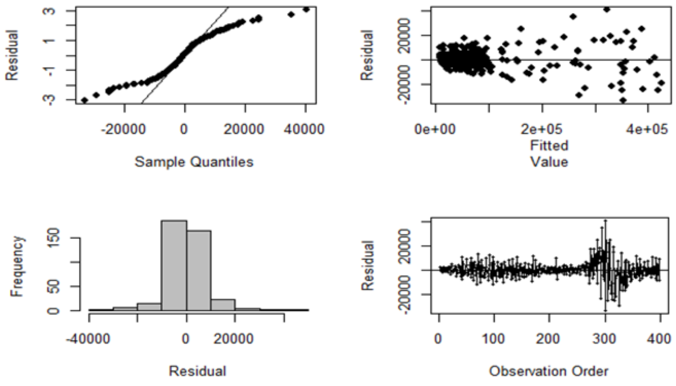

Figure 14 Residuals plots for daily confirmed cases.

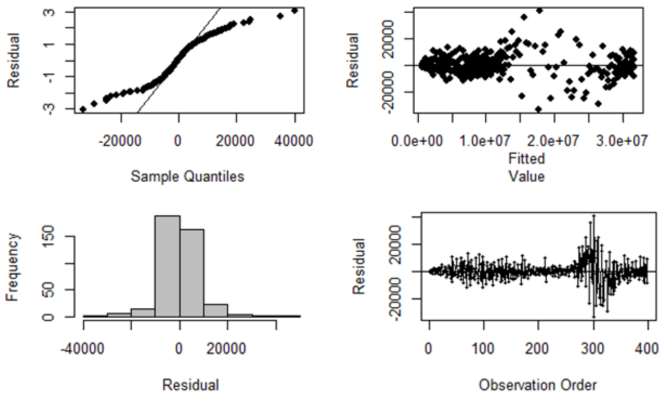

Figure 15 Residuals plots for total confirmed cases.

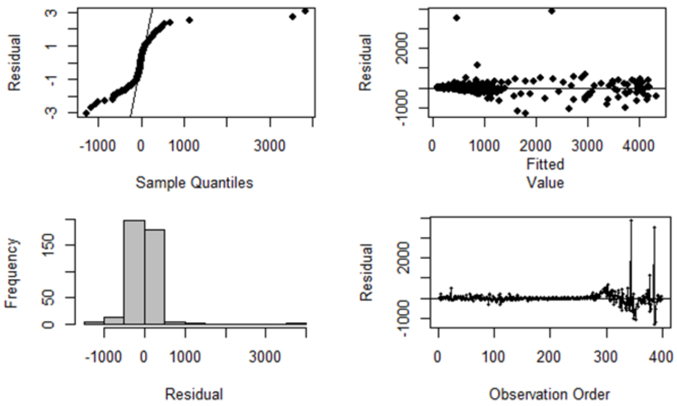

Figure 16 Residuals plots for daily decease cases.

Table 5 Durbin Watson Test and accuracy of forecast of models for three cases

Cases	Durbin Watson test	Accuracy of forecast		
		ME	MASE	ACF
Daily confirmed	1.98	-0.82801	0.6824	-0.06764
Total confirmed	1.98	-1.47696	0.682913	-0.06631
Daily deceased	2	-0.27882	0.861764	0.000392

Conclusions

It has been found that ARIMA (3,1,2), ARIMA(3,2,2) and ARIMA(0,1,2) are appropriate models for daily confirmed, total confirmed and daily deceased cases of COVID-19 respectively. The forecast for next 120days (from August 4, 2021 to November, 2021) have been calculated and presented graphically. PP and KPSS test have been applied for stationarity for the selected three cases of COVID-19. ACF and PACF plots as well as values of AIC and BIC of models have been used to select the appropriate models. Accuracy of forecast models has been checked using Mean Error (ME), Mean Absolute Standard Error (MASE) and Auto Correlation Functions (ACF). According to the point estimation, there may be very less chance to increase the COVID-19 cases up to November 2021. It was found declining pattern for daily confirmed cases as well daily deceased cases. It is observed from the analysis as well interval estimation of the forecast that there are fewer chances to increase

COVID-19 cases up to November, 2021. We may conclude that the situation may improve in coming days in India because almost 20 percent people of India have been vaccinated and still the vaccination process is going on with full awareness. It may increase; if any other novel COVID-19 cases start in near future otherwise situation may improve.

Acknowledgments

The authors are grateful to the chief editor as well anonymous referee for giving valuable comments to improve the quality of the paper.

Conflicts of interest

Author declares there are no conflicts of interest.

Funding

None.

References

1. COVID-19. In: COVID19 INDIA. (Accessed 2020).
2. Liu Z, Magal P, Seydi O, Webb G. A COVID-19 epidemic model with latency period [published online ahead of print, 2020 Apr 28]. *Infect Dis Model*. 2020;323–337.
3. WHO: COVID-19 Coronavirus pandemic. (Accessed 2020).
4. Hamzah F A, Lau C, Nazri H, et al. Corona Tracker: worldwide COVID-19 outbreak data analysis and prediction. *Bull World Health Organ.* 2020;1:32.
5. Shakil M, Munim Z, Tasnia M, et al. COVID-19 and the environment: A critical review and research agenda. *Science of the Total Environment.* 2020;745:141022.
6. Acuña-Zegarra M, Santana-Cibrian M, Velasco-Hernandez J. Modeling behavioral change and COVID-19 containment in Mexico: A trade-off between lockdown and compliance. *Mathematical Biosciences*. 2020;325:108370.
7. Singh R, Adhikari R. Age-structured impact of social distancing on the COVID-19 epidemic in India. *arXiv preprint arXiv*. 2020;2003.12055:1–9.
8. Wagh C, Mahalle P, Wagh S. Epidemic peak for COVID-19 in india, 2020. *Preprints*. 2020;1–7.
9. De Livera AM, Hyndman RJ, Snyder RD. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association. 2011;*106(496): 1513-1527.
10. Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. Chaos, *Solitons and Fractals*. 2020;135:1–10.
11. Batista M. Estimation of the final size of the COVID-19 epidemic. MedRxiv. 2020.
12. Benvenuto D, Giovanetti M, Vassallo L, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset, *Data in Brief.* 2020;29:105340.
13. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France, *Chaos, Solitons & Fractals*. 2020;134:109761.
14. Giordano G, Blanchini F, Bruno R. et al. A SIDARTHE model of COVID-19 epidemic in italy. ArXiv preprint. *ArXiv*. 2020;2003:0986.
15. Gupta R, Pal SK. Trend Analysis and Forecasting of COVID-19 outbreak in India. *MedRxiv.* 2020.
16. Kumar P, Kalita H, Patairiya S, et al. Forecasting the dynamics of COVID-19 Pandemic in Top 15 countries in April 2020 through ARIMA Model with Machine Learning Approach. *MedRxiv*. 2020.
17. Read JM, Bridgen JR, Cummings D A, et al. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. *MedRxiv*. 2020.
18. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, *The Lancet.* 2020;395(10225):689–697.
19. Zhao S, Lin Q, Ran J, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak, *International Journal of Infectious Diseases*. 2020; 92:214–217.
20. Perron P. Trends and Random Walks in Macroeconomic Time Series. *Journal of Economic Dynamics and Control*. 1988;12(2–3):297–332.
21. Kwiatkowski D, Phillips PCB, Schmidt P, et al. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*. 1992;54:159–178.