Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 223-238

Analysis and Forecasting of COVID-19 Cases Across Hotspot States of India

Khimya Tinani, K. Muralidharan, Akash Deshmukh, Bhagyashree Patil, Tanvi Salat and Rajeshwari Rajodia
Department of Statistics, Faculty of Science,
The Maharaja Sayajirao University of Baroda, Vadodara 390002, India.

Received: 18 May 2020; Revised: 15 June 2020; Accepted: 16 June 2020

Abstract

This paper attempts to develop a model to predict Novel Coronavirus affected cases in India. The virus is officially named as SARS-CoV-2 and was declared as a pandemic by WHO on 11th March 2020. This pandemic erupted in the Wuhan city of the People's Republic of China in December 2019. By now the whole world is in the grip of this virus. The first case of the COVID-19 in India was reported on 30th January 2020 in the state of Kerala. In India, the Ministry of Health and Family Welfare (MOHFW) keeps the track of COVID-19 cases daily. As of 14th June 2020, the total number of confirmed, recovered, and death cases in India are 332424, 169798 and 9520 respectively. The corresponding world statistics are 7900924, 3769712 and 433065 respectively. The disease is infectious and contagious and is affecting the health of people at large. The government and administration are trying hard to control the disease, and trying to find an effective treatment. This research aims to forecast the number of confirmed cases, recoveries and deaths of India and its six hotspot states (Maharashtra, Delhi, Tamil Nadu, Madhya Pradesh, Rajasthan, and Gujarat). To check the accuracy of the model, the first round of forecast is done from 15/4/2020 to 25/04/2020 based on the data available from 30th January 2020 to 14th April 2020. The second round of forecast is done from 16/05/2020 to 30/06/2020 based on the actual data from 30/01/2020 to 15/05/2020. Auto-Regressive Integrated Moving Average (ARIMA) model has been used to forecast the trend of COVID-19 cases in R programming.

Key words: COVID-19; Coronavirus; ARIMA; Forecast; Pandemic; Epidemic.

1. Introduction

Coronaviruses are commonly found in humans and animals. COVID-19 is an acronym that stands for the coronavirus disease of 2019. Common symptoms include fever, body ache, tiredness, and difficulty in breathing. Many affected people do not show any symptoms. The virus spreads within populations via respiratory droplets and close contact. Symptoms usually start 4 days after a person is infected with the virus. But in some people, it can take even longer for symptoms to appear or an infected person gets recovered without the appearance of any symptoms. The death rate of patients affected with COVID-19 is very less. The risk of becoming severely sick from COVID-19 increases with age. People who are critically ill are more prone to death if affected by COVID-19. The medicine for the treatment of COVID-19 is not found and the vaccine for COVID-19 is not available till 14th April 2020. However, the studies are being conducted by different countries. Since this is a statistical modeling-based

Corresponding Author: Khimya Tinani
E-mail: khimya27@yahoo.com

study, we deliberately avoid any detailed descriptions about the virus and its genesis. But to understand the inference part of this analysis, we need to supplement some information regarding its transmission and spread. The COVID-19 has four stages of transmission in line with other infectious diseases. In stage-1 the first appearance of the disease is through people with travel history, with everyone contained, their sources can be traced, and no local spread from those affected. The number of those infected would be quite low at this stage. Stage-2 is the local transmission when those who were infected and have a travel history spread the virus to close friends or family. At this stage, every person who comes in contact with the infected can be traced and isolated. Stage-3 is the community transmission when infections happen in public and a source for the virus cannot be traced. At this stage, large geographical lockdowns become important as random members of the community start developing the disease. Stage-4 is when the disease becomes an epidemic in a country, such as it was in China, with large numbers of infected people and the growing number of deaths with no end in sight. The World Health Organization declared it a pandemic. In the absence of a vaccine, social distancing has emerged as the most widely adopted strategy for mitigating and control of the virus. In India, the first novel coronavirus infection was reported on January 30 at Kerala. The cases increased to three by February 3. After this, no new cases were reported until March 1. On March 2, India reported two more positive cases, one each from Delhi and Hyderabad. By March 15, the total number of confirmed patients reached 107, most of which were linked to people with the travel history to affected countries and since then, the number of positive cases is continuously increasing. India observed a 14-hour voluntary public curfew on 22nd March 2020. This was followed by a nationwide lockdown for 21 days starting from 24 March 2020 and later extended to 3 May 2020, as the cases affected and deaths are increasing. The Indian Government feels that in the absence of lockdown this contagious disease may spread to a greater number of people and the number of hospitals may turn to be insufficient with limited equipment for the treatment of Covid-19 cases. However, understanding the seriousness of the issue, we feel that, constructing a good statistical model for inference and forecasting is the best we can contribute to this current subject. If the model fits well, then an estimate of the need for healthcare infrastructure, investment, and manpower can be anticipated.

In this paper, based on the data from January 30, 2020, till April 14, 2020, the first round of forecast was done day-wise for 11 days: 15/04/2020 till 25/04/2020 and the accuracy of the model was checked. The second round of forecast is done for 46 days: 16/05/2020 till 30/06/2020 based on actual data from January 30, 2020, till May 15, 2020. Since the forecasts for the number of days in the second round are more, we have presented only the weekly figures in the table. Auto-Regressive Integrated Moving Average (ARIMA) model has been used to predict the trend of COVID-19 cases using R programming.

2. Review of Literature

Petropoulos and Makridakis (March 2020) published the research article on forecasting the novel coronavirus COVID-19. Their paper describes the timeline of a live forecasting exercise with massive potential implications for planning and decision making and provides forecasts for the confirmed cases of COVID-19. Their study focuses on the cumulative daily figures aggregated globally of the three main variables like confirmed cases, deaths and recoveries. In their forecast, they predicted the cases for three variables in the period of 5 rounds. Kai Liu *et al.* (March 2020) studied that the mortality of elderly patients with COVID-19 is higher than that of young and middle-aged patients and elderly patients with COVID-19 are more likely to progress to severe disease. Khot and Nadkar (March 2020) published a valuable research paper on “The 2019 Novel Coronavirus Outbreak-A Global Threat’. They had shown new insights into the pathophysiology, transmission dynamics, clinical features and management of this virus are developing. They said it is a highly transmissible infection but mortality is less compared to SARS and MERS. National and International health care agencies have shown appropriate co-ordination in the handling of this outbreak up till now and further international cooperation is the need of the hour. Lina *et al.* (March 2020) published a research paper on "A conceptual model for the coronavirus disease 2019 outbreak in Wuhan, China with individual reaction and governmental action". In this paper, their main purpose was to propose a conceptual model to address the individual reaction and governmental action, as well as the time-varying reporting rate. Schueller *et al.* (April 2020) had done research on COVID-19 in India on the potential impact of the 21-day Lockdown which was announced with effect from 25 March 2020 and other long-term policies. This lockdown is expected to avert a sudden and large increase in the number of infections in the short term. Additionally, interventions such as social distancing and isolation of infected individuals over several months could reduce peak infections and also interventions such as frequent hand washing, reduced mass gatherings, contact tracing, and quarantines could slow transmission and reduce overall infections. Read *et al*. (January 2020) studied and show the important information for the crisis management against the novel Coronavirus, early estimation of epidemiological parameters and epidemic predictions. Also, researchers proved that the SIR-family models at different complex levels can well capture the basic mechanism of the epidemic transmission. Liu *et al*. (February 2020) discussed on the reproductive number of COVID-19 is higher compared to SARS Coronavirus. They reviewed the basic reproduction number of the COVID-19 virus. Reproduction number is an indication of the transmissibility of a virus, representing the average number of new infections generated by an infectious person in a population. Khrapov and Loginova (2020) presented a research paper on mathematical modeling of coronavirus COVID-19, the authors used a modified system of differential equations constructed according to the SIR compartmental model. The optimal values of the model parameters, that describe the statistical data precisely, were found. Miller *et al*. (2020) published their study with an emphasis on the correlation between universal BCG vaccination policy and how it reduced morbidity and mortality of COVID-19 patients. They also found that countries without universal policies of BCG vaccination (Italy, Nederland, USA are some of them) have been more severely affected compared to countries with universal and long-standing BCG policies. BCG vaccination is a potential new tool in the fight against COVID-19. Probably a detailed statistical and mathematical treatment of modeling on this virus was done by Lin *et al.* (2019). For mathematical treatment, they used infectious disease prediction models based on differential equation prediction models and time series prediction models based on statistics and random processes. They also used the internet-based infectious disease prediction model and machine learning methods to substantiate the findings. Tania *et al*. (2020) published the research paper on "Forecasting of COVID-19 confirmed cases in different countries with ARIMA models". The aim of this study was first to find the best prediction models for daily confirmed cases in countries with a high number of confirmed cases in the world and second to predict confirmed cases with these models in order to have more readiness in healthcare systems. Ribeiro *et al.* (2020) developed efficient short-term forecasting models for forecasting the number of future cases. In their paper, they are using an autoregressive integrated moving average (ARIMA), cubist regression (CUBIST), random forest (RF), ridge regression (RIDGE), support vector regression (SVR) and stacking-ensemble learning models for evaluating in the task of time series forecasting with one, three, and six-days ahead the COVID-19 cumulative confirmed cases in ten Brazilian states with a high daily incidence. The models' effectiveness is evaluated based on the improvement index, mean absolute error, and symmetric mean absolute percentage error criteria. The ranking of models, from the best to the worst regarding the accuracy, in all scenarios, is SVR, stacking-ensemble learning, ARIMA, CUBIST, RIDGE, and RF models.

3. Objectives

Forecast the number of COVID-19 confirmed cases for India as well as across the six hotspot states of India. Also, predict the number of deaths and recoveries amongst the number of cases of COVID-19 of India and across the hotspot states of India.

4. Data Source

This study has been conducted based on daily confirmed cases, deaths and recoveries of COVID-19 of India and only those states that are considered as hotspots of India. The data was collected from the official Indian website of COVID-19: https://www.mohfw.gov.in/ from 30 January 2020 to 15 May 2020.

5. Data Visualization

Table 1: Mortality rate and Recovery rate of six hotspots states of India for the period 30/01/2020 to 25/04/2020

Hotspots States of India	Confirmed cases	Death cases	Recovered cases	Mortality Rate per thousand	Recovery Rate per thousand
Maharashtra	7628	322	1076	42.2129	141.0593
Gujarat	3071	133	282	43.3083	91.82677
Delhi	2625	54	869	20.5714	331.0476
Rajasthan	2083	34	513	16.3226	246.2794
Madhya Pradesh	1945	100	281	51.4138	144.473
Tamil Nadu	1821	23	960	12.6304	527.1829

From the above table, it is observed that the mortality rate in Madhya Pradesh is highest when compared with other hotspots states of India. While the recovery rate in Tamil Nadu is highest and on other side mortality rate is minimal compared to other hotspot states of India.

6. Analysis and Forecasting

6.1. ARIMA Model

Autoregressive Integrated Moving Average (ARIMA) is a stochastic approach of modeling which can be used for calculating the probability of a future value lying in a specified interval of limits. It consists of two models Autoregressive Process (AR) and Moving Average Process (MA) bind together by (I) the integration part. ARIMA models are generally used to analyze time series data for better understanding and forecasting. The ARIMA model is denoted as ARIMA (p, d, q), where the parameter p refers to the order of the AR process, q refers to the order of the MA process, and d refers to the order of differencing it takes to make the series stationary. In this study, the ARIMA model has been developed to forecast the confirmed cases, death cases and recovered cases of India cumulatively and its six hotspot states.

The ARIMA model for Confirmed cases is given as:

$$X_t = \Lambda + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots\ldots + \alpha_p X_{t-p} + \varepsilon_t + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \ldots\ldots + \Phi_q \varepsilon_{t-q} \qquad (1)$$

where, X_t shows the forecasted values of confirmed cases, Λ is the intercept term, also estimated by the model, X_{t-i} is the lag variable at the time $t-i$ of the series, i=1, 2, …, p, α_i is the coefficient of AR process that the model estimates, ε_t is the error term and Φ_j is the coefficient of MA process where, j=1, 2, …, q.

The ARIMA model for Death cases is given as:

$$\Upsilon_t = \psi + \beta_1 \Upsilon_{t-1} + \beta_2 \Upsilon_{t-2} + \ldots\ldots + \beta_p \Upsilon_{t-p} + \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \Theta_2 \varepsilon_{t-2} + \ldots\ldots + \Theta_q \varepsilon_{t-q} \qquad (2)$$

where, Υ_t shows the forecasted values of death cases, ψ is the intercept term, also estimated by the model, Υ_{t-i} is the lag variable at the time $t-i$ of the series, i=1, 2, …, p, β_i is the coefficient of AR process that the model estimates, ε_t is the error term and Θ_j is the coefficient of MA process where, j = 1, 2, …, q.

The ARIMA model for Recovered cases is given as:

$$Z_t = \zeta + \gamma_1 Z_{t-1} + \gamma_2 Z_{t-2} + \ldots\ldots + \gamma_p Z_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots\ldots + \phi_q \varepsilon_{t-q} \qquad (3)$$

where, Z_t shows the forecasted values of recovered cases, ζ is the intercept term, also estimated by the model, Z_{t-i} is the lag variable at the time $t-i$ of the series, i=1, 2, …, p, γ_i is the coefficient of AR process that the model estimates, ε_t is the error term and ϕ_j is the coefficient of MA process where, j =1, 2, …, q.

The first step to build an ARIMA model is to make the time series stationary. So, to make a series stationary, the most common approach is to difference it. Augmented Dickey Fuller test (ADF test) is a common statistical test used to test whether a given time series is stationary or not. The null hypothesis assumes that the series is non-stationary. ADF test is fundamentally a statistical significance test. That means, there is a hypothesis testing involved with a null and alternative hypothesis and as a result, a test statistic is computed and p-values get reported. It is from the test statistic and the p-value, we can make an inference as to whether a given series is stationary or not. For the identification of the model, the task is to find out the appropriate values of p and q with the help of autocorrelation function (ACF) and partial autocorrelation function (PACF) graph values. The initial number of the ARIMA model was guessed through the autocorrelation function (ACF) graph and partial autocorrelation (PACF) graph. ACF plot is merely a bar chart of the coefficients of correlation between a time series and lags of itself. The PACF plot is a plot of the partial correlation coefficients between the series and lags of itself. According to these plots, the p and q parameters of ARIMA models were guessed. Then the guess models were compared according to Akaike Information Criterion (AIC) value, treating minimum as the best. The reason for choosing AIC is because of its wide acceptance as a statistical measure model selection. It is used to quantify the goodness of fit of the model. When comparing two or more models, the one with the lowest AIC is generally considered to be closer to real data. The appropriate ARIMA model then identified for the particular datasets and the parameters are estimated accordingly.

Having chosen the specific ARIMA model and its parameters estimated, the next step is to carry out a diagnostic check to see whether the model fits the data completely well. That is done by checking the residuals estimated from this model which are termed as white noise error or pure random error. This will decide if the chosen model fits the data well or not. For this, we use the Ljung-Box test introduced in (1978) which as a diagnostic tool to test the lack of fit of a time series model. The null hypothesis of the Ljung-Box test is given by H_0: The model does not show a lack of fit and the alternative hypothesis is H_1: the model does show a lack of fit. For a time series Y of length n, the Ljung-Box test statistic is defined as:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{r}_k}{n-k} \tag{4}$$

where $\hat{r}_k$ is the estimated autocorrelation of the series at lag k, and m is the number of lags being tested with a significant level α. We reject the null hypothesis and say that the model has significant lack of fit if $Q > {\chi^2}_{1-\alpha,h}$ where ${\chi^2}_{1-\alpha,h}$ is the chi-square distribution table value with h degrees of freedom and significant level α. Because the test is applied to residuals, the degrees of freedom must account for the estimated model parameters so that $h = m–p–q$, where p and q indicate the number of parameters from the ARIMA (p, d, q) model fit to the data. In Statistical package R, the Ljung-Box test can be run with the help of *Box.test* function.

After prediction, the accuracy is measured in percentage. We have used the Mean Absolute Error (MAE) method to compute the accuracy. Firstly, the predicted values and the actual values are stored in a single matrix with two columns, namely predicted value and actual value respectively. Then the error between the 2 columns is computed where, error =|actual value –predicted value|. The accuracy is calculated by,

$$Accuracy = 1 - \frac{error}{(actual\ value)} \tag{5}$$

which is generally reported in percentages.

6.2. First round of forecasts for the period: 15/04/2020 to 25/04/2020

Our focus is on the cumulative daily figures aggregated for India over the period from January 30, 2020 till April 14, 2020. While the data patterns show an exponential increase, the trend of confirmed cases comes to hold after it first entered India on 30th January 2020 until February 2020. From March beginning there was a sudden increase in cases, while deaths started to happen frequently only after 11th March onwards. The recovery of patients started to happen simultaneously from mid-February onwards. We have aimed our research to forecast the number of confirmed cases, recoveries and deaths of India and its six hotspot states (Maharashtra, Delhi, Tamil Nadu, Madhya Pradesh, Rajasthan and Gujarat). Based on the data from January 30, 2020, till April 14, 2020, the first round of forecast was done day-wise for the period of 11 days: 15/04/2020 to 25/04/2020 and the accuracy of the model was checked. The analysis is done in R programming and the necessary packages: library(hrbrthemes), library(dplyr), library(ggplot2), library(tseries), library(forecast) are loaded.

Now before we analyze the time series data for actual forecast, we use the Augmented Dickey Fuller test to check the stationary of the time series observations. The null hypothesis (H_0) for the test is that the data is not stationary whereas the alternative hypothesis (H_1) is that the data is stationary. The level of significance is taken to be 0.05. The output is obtained for confirmed cases using *adf.test* function in R programming. Here, the p-value turns out to be

0.99. We thus fail to reject our H_0 and conclude that the data is not stationary. We now have to work on the stationarity of the data. After differencing the time series for consecutively for two times, the *p*-value is obtained as 0.01, which is less than 0.05, and hence we reject the null hypothesis and conclude that the time series for confirmed cases is stationary. Since the order of differencing is 2, $d = 2$. Similarly, we have found that stationary time series for deaths and recoveries cases.

Figures 1 and 2 show the ACF and PACF plots for confirmed cases. These plots are used for choosing the model parameters for confirmed cases. Similarly, we have found model parameters for deaths and recoveries using ACF and PACF plots.

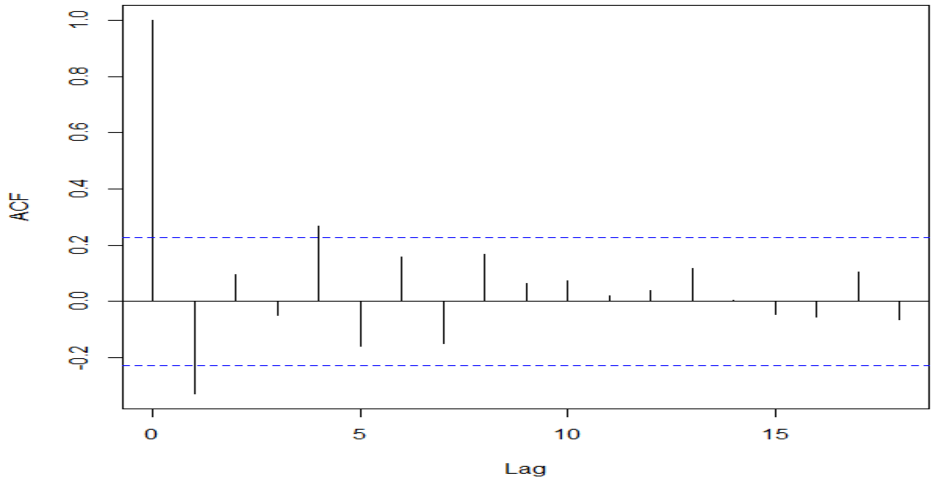

Figure 1: Plot of ACF for confirmed cases

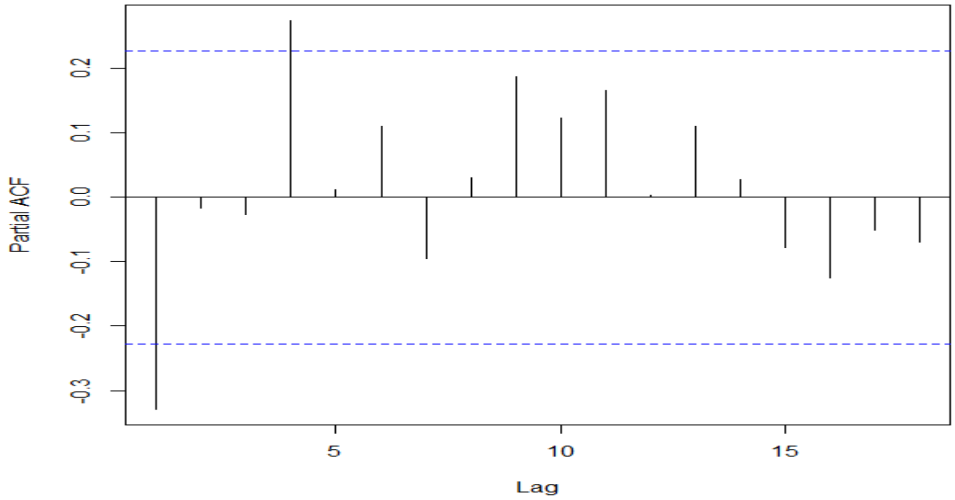

Figure 2: Plot of PACF for confirmed cases

According to ACF and PACF plots, the *p* and *q* parameters of ARIMA models are guessed. These guess models are compared according to AIC value. Table 2 presents all those ARIMA models with corresponding AIC values for all three types of cases.

Table 2: ARIMA models with all possible values of AIC for India

Confirmed	AIC	Deaths	AIC	Recovered	AIC
ARIMA(0,2,0)	853.724	ARIMA(0,2,0)	473.335	ARIMA(1,2,0)	627.8381
ARIMA(2,2,0)	848.897	ARIMA(1,2,2)	399.478	ARIMA(5,2,0)	599.7987
ARIMA(3,2,0)	850.859	ARIMA(1,2,1)	423.039	ARIMA(1,2,2)	617.8326
ARIMA(1,2,0)	849.184	ARIMA(1,2,0)	425.797	ARIMA(2,2,0)	606.2791
ARIMA(0,2,1)	850.707	ARIMA(0,2,1)	439.349	ARIMA(0,2,0)	627.6912
ARIMA(0,2,0)	853.724	ARIMA(0,2,0)	473.335	ARIMA(1,2,0)	627.8381

The model which has the least AIC is selected as the best model. Accordingly, the best ARIMA models for forecasting the number of daily confirmed, deaths and recovered cases for India are ARIMA(2,2,0), ARIMA(1,2,2), ARIMA(5,2,0) respectively for India. The first round of forecast is shown in figure 3. The same in actual numbers are presented in Table 3.

The equation corresponding to the best ARIMA(2,2,0) model for confirmed cases is given by

$$X_t = 15.3463 - 0.3524X_{t-1} + 50.1764X_{t-2} + \varepsilon_t \qquad (6)$$

The equation corresponding to the best ARIMA(1,2,2) model for death cases is given by

$$Y_t = 0.4077 - 0.2613Y_{t-1} + \varepsilon_t - 0.7937\varepsilon_{t-1} + 0.7014\varepsilon_{t-2} \qquad (7)$$

The equation corresponding to the best ARIMA(5,2,0) model for recovery cases is given by,

$$Z_t = 4.860 + 0.085Z_{t-1} + 0.261Z_{t-2} + 0.444Z_{t-3} + 0.632Z_{t-4} + 0.622Z_{t-5} + \varepsilon_t \qquad (8)$$

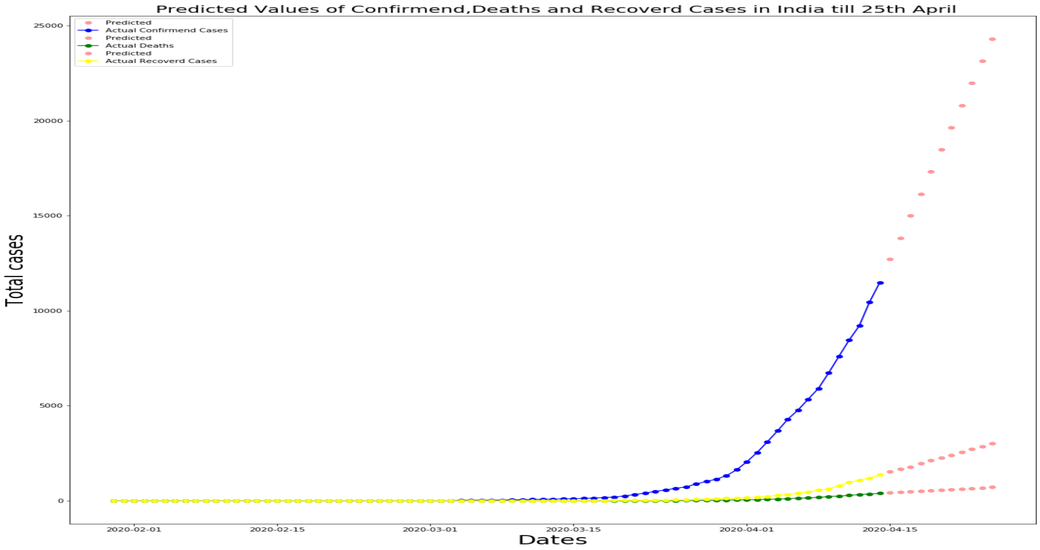

Figure 3: Plot of actual and forecasts of COVID-19 cases in India

The blue dots represent the actual confirmed cases, yellow dots represent recovered cases and green dots represent the actual deaths. The extended red dots represent forecasted COVID-19 cases.

Table 3: Actual and forecast values of COVID-19 with 95% CI for India

Date	Actual values			Forecast values		
	Confirmed	Death	Recovered	Confirmed	Death	Recovered
15-04-20	12370	422	1509	12707 (12564, 12850)	427 (420, 433)	1538 (1512, 1563)
16-04-20	13431	448	1767	13817 (13534, 14100)	456 (446, 465)	1664 (1613, 1715)
17-04-20	14353	486	2040	15012 (14525, 15498)	485 (469, 500)	1793 (1719, 1866)
18-04-20	15724	521	2466	16152 (15438, 16865)	514 (490, 537)	1965 (1861, 2069)
19-04-20	17304	559	2854	17330 (16352, 18308)	543 (510, 576)	2130 (1983, 2278)
20-04-20	18543	592	3273	18482 (17217, 19748)	572 (528, 615)	2261 (2065, 2458)
21-04-20	20080	645	3976	19653 (18072, 21234)	601 (546, 656)	2404 (2158, 2649)
22-04-20	21372	681	4370	20811 (18894, 22728)	630 (562, 698)	2575 (2273, 2877)
23-04-20	23039	721	5012	21978 (19702, 24253)	659 (578, 740)	2735 (2367, 3102)
24-04-20	24447	780	5496	23138 (20485, 25792)	688 (593, 783)	2871 (2434, 3308)
25-04-20	26282	824	5939	24303 (21253, 27354)	717 (607, 827)	3021 (2512, 3530)

From the above table, it is noted that the day-wise estimated figures for confirmed cases from 15 April 2020 to 25 April 2020 are nearly the same. However, the day-wise estimated recoveries are less than the actual values. To estimate model adequacy, the Ljung-Box test which is a diagnostic tool is used to test the lack of fit of a time series model. The output is obtained by using the *Box.test* function in R programming. The null hypothesis, H_0: The model does not show a lack of fit. The alternative hypothesis, H_1: The model does show a lack of fit. Here, for confirmed cases *p*-value is 0.7315, for deaths *p*-value is 0.49863 and for recoveries, the *p*-value is 0.9585. As for all the cases, *p*-value is greater than 0.05, hence we do not reject the null hypothesis and conclude that our model does not show a lack of fit. The accuracy of prediction for India is computed by averaging the accuracies obtained by the algorithm of ARIMA modeling. As per this modeling, the accuracy for confirmed cases is 98%, for the deaths 97% and for the recoveries is 78%.

Now we will forecast the figures for the highly affected states in India assuring that the data is stationary and reliable to forecast. The final models that are reported in table 4 have the lowest AIC values for all hotspot states of India. The equations of best ARIMA model can be mentioned for all the hotspots states of India in the same way as we mentioned for India. To estimate model adequacy, the Ljung-Box test which is a diagnostic tool is used to test the lack of fit of a time series model. The outputs for all six hotspot states of India are given in table 4. The null hypothesis, H_0: The model does not show a lack of fit. The alternative hypothesis, H_1: The model does show a lack of fit. The *p*-value for state Rajasthan is less than 0.05 for deaths and recoveries, hence we reject the null hypothesis and conclude that model does show lack of fit whereas *p*-value for Rajasthan is more than 0.05 for confirmed cases, hence we do not reject the null hypothesis and conclude that model does not show lack of fit for confirmed cases. For the other hotspot states p-value is greater than 0.05 for confirmed, deaths and recovery cases, hence we do not reject the null hypothesis and conclude that our model does not show a lack of fit.

Table 4: The best ARIMA models with least AIC for six hotspots states of India

Hotspot States	Cases	ARIMA Model	AIC	Ljung-Box p-value
Maharashtra	Confirmed	ARIMA(1,2,0)	695.1979	0.9091
	Deaths	ARIMA(2,2,2)	333.7695	0.9026
	Recovered	ARIMA(1,2,1)	519.3081	0.9072
Delhi	Confirmed	ARIMA(2,2,0)	741.8226	0.3344
	Deaths	ARIMA(3,2,0)	151.5864	0.7458
	Recovered	ARIMA(0,2,2)	233.4595	0.8812
Madhya Pradesh	Confirmed	ARIMA(2,2,0)	565.2183	0.1515
	Deaths	ARIMA(1,2,1)	210.2508	0.9744
	Recovered	ARIMA(0,2,1)	322.8164	0.1021
Tamil Nadu	Confirmed	ARIMA(2,2,2)	614.3582	0.1434
	Deaths	ARIMA(0,2,2)	57.20068	0.9963
	Recovered	ARIMA(0,2,5)	344.3379	0.8514
Gujarat	Confirmed	ARIMA(3,2,1)	557.0561	0.9976
	Deaths	ARIMA(3,2,0)	74.23162	0.9909
	Recovered	ARIMA(0,2,2)	297.4889	0.9622
Rajasthan	Confirmed	ARIMA(0,2,3)	551.4806	0.6519
	Deaths	ARIMA(0,2,1)	168.3248	0.0320
	Recovered	ARIMA(2,2,1)	496.3972	0.0283

Forecast values of ARIMA models with a confidence interval for six hotspot states of India are given in Table 5.

Table 5: Forecast values of COVID-19 cases with 95% CI for six hotspot states of India

Date	Cases	Maharashtra	Delhi	Madhya Pradesh	Tamil Nadu	Gujarat	Rajasthan
15-04-20	Confirmed	3028 (2977, 3079)	1801 (1732, 1869)	764 (744, 785)	1255 (1226, 1283)	735 (715, 754)	1093 (1074, 1112)
	Death	191 (187, 196)	32 (31, 34)	59 (57, 61)	12 (12, 13)	30 (29, 31)	11 (10, 13)
	Recovered	277 (262, 293)	32 (30, 35)	72 (68, 76)	85 (81, 90)	64 (61, 68)	163 (150, 176)
16-04-20	Confirmed	3375 (3275, 3475)	1998 (1906, 2090)	842 (816, 869)	1318 (1263,1373)	803 (758, 849)	1195 (1155, 1234)
	Death	206 (200, 212)	35 (33, 38)	63 (60, 66)	13 (12, 14)	32 (31, 33)	12 (10, 14)
	Recovered	299 (276, 322)	34 (31, 37)	81 (74, 88)	94 (87,101)	70 (66, 75)	172 (154, 189)
17-04-20	Confirmed	3723 (3561, 3884)	2140 (2003, 2276)	947 (911, 983)	1367 (1278, 1455)	868 (797, 939)	1291 (1221, 1361)
	Death	222 (211, 232)	39 (35, 43)	68 (64, 73)	14 (12, 16)	34 (33, 36)	13 (10, 16)
	Recovered	320 (288, 353)	35 (31, 39)	89 (79, 99)	102 (94, 111)	76 (70, 83)	190 (168, 212)

18-04-20	Confirmed	4070 (3839, 4301)	2366 (2178, 2555)	974 (921, 1028)	1404 (1272, 1537)	937 (843, 1032)	1387 (1285, 1489)
	Death	237 (221, 252)	42 (37, 47)	73 (67, 79)	15 (13, 17)	36 (34, 38)	14 (10, 17)
	Recovered	342 (299, 384)	37 (32, 42)	98 (85, 111)	116 (105, 126)	82 (74, 91)	206 (177, 236)
19-04-20	Confirmed	4417 (4109, 4726)	2530 (2292, 2767)	1068 (1003, 1133)	1454 (1271, 1637)	1014 (893, 1135)	1483 (1347, 1619)
	Death	252 (231, 272)	45 (39, 52)	77 (70, 85)	16 (13, 18)	39 (36, 41)	15 (11, 19)
	Recovered	363 (309, 417)	38 (32, 44)	106 (90, 123)	126 (112, 140)	88 (78, 99)	219 (183, 256)
20-04-20	Confirmed	4765 (4371, 5158)	2716 (2418, 3015)	1152 (1072, 1232)	1512 (1277, 1748)	1086 (935, 1238)	1579 (1407, 1752)
	Death	267 (240, 294)	48 (40, 56)	82 (72, 92)	16 (13, 20)	41 (38, 44)	16 (11, 21)
	Recovered	384 (318, 450)	39 (32, 47)	115 (95, 135)	137 (118, 155)	94 (82, 107)	235 (192, 279)
21-04-20	Confirmed	5112 (4628, 5596)	2912 (2551, 3272)	1189 (1090, 1288)	1562 (1271, 1852)	1157 (972, 1342)	1676 (1464, 1887)
	Death	282 (249, 315)	51 (42, 61)	87 (75, 98)	17 (13, 21)	43 (39, 47)	17 (11, 22)
	Recovered	405 (327, 483)	41 (32, 50)	123 (100, 147)	147 (125, 170)	100 (85, 115)	251 (199, 303)
22-04-20	Confirmed	5459 (4878, 6041)	3085 (2658, 3511)	1289 (1175, 1403)	1603 (1252, 1955)	1226 (1006, 1446)	1772 (1519, 2024)
	Death	297 (257, 337)	55 (43, 66)	91 (77, 105)	18 (14, 22)	45 (41, 50)	17 (11, 24)
	Recovered	426 (335, 518)	42 (32, 52)	132 (104, 160)	158 (130, 186)	106 (89, 124)	265 (205, 326)
23-04-20	Confirmed	5807 (5122, 6491)	3278 (2780, 3775)	1357 (1224, 1491)	1652 (1235, 2070)	1299 (1043, 1555)	1868 (1572, 2164)
	Death	312 (265, 360)	58 (44, 71)	96 (80, 112)	19 (14, 23)	46 (42, 53)	18 (11, 25)
	Recovered	447 (342, 553)	44 (32, 55)	140 (109, 172)	169 (136, 201)	112 (92, 132)	281 (211, 350)
24-04-20	Confirmed	6154 (5362, 6947)	3461 (2890, 4031)	1408 (1254, 1562)	1708 (1222, 2193)	1371 (1077, 1665)	1964 (1623, 2305)
	Death	328 (272, 383)	61 (45, 76)	101 (82, 119)	19 (14, 25)	50 (44, 56)	19 (12, 27)
	Recovered	469 (349, 589)	45 (32, 58)	149 (113, 185)	179 (141, 217)	118 (96, 141)	296 (217, 375)
25-04-20	Confirmed	6502 (5595,7408)	3644 (2996, 4292)	1506 (1334, 1678)	1757 (1201, 2313)	1442 (1108, 1777)	2060 (1672, 2449)
	Death	343 (279, 406)	64 (46, 82)	105 (84, 126)	20 (14, 26)	52 (45, 58)	20 (12, 28)
	Recovered	490 (355, 625)	47 (32, 61)	158 (117, 198)	190 (146, 233)	124 (99, 150)	311 (222, 400)

From the above table, it can be noted that the day-wise estimated figures for confirmed cases from 15 April 2020 to 25 April 2020 are nearly the same. However, the day-wise estimated recoveries and deaths are less than the actual values. The accuracy of prediction for six hotspots states of India is computed by averaging the accuracies obtained by the algorithm of ARIMA modeling. The result is given below in Table 6.

Table: 6 Model Accuracy for six hotspot states of India

Hotspot States	Confirmed	Deaths	Recovered
Maharashtra	92%	95%	71%
Delhi	68%	91%	66%
Madhya Pradesh	81%	80%	79%
Tamil Nadu	98%	85%	53%
Rajasthan	76%	26%	31%
Gujarat	82%	77%	78%

For all six hotspots states of India, the ARIMA model accuracy of confirmed cases forecasted is 83% on an average, which indicates that ARIMA gives good accuracy of prediction. On the other hand, model accuracy for death and recovery cases in six hotspots states of India is around 76% and 58% respectively. This seems we need a better model for forecasting death and recovery cases in hotspots states of India.

6.3. Second round of forecasts for the period: 16/05/2020 till 30/06/2020

While writing this paper, the number of cases in India is doubling up every day, and hence the prediction after 25th April may not match with our estimated values. The forecast of COVID-19 cases until 25thApril 2020 is nearly the same as per our actual cases. This needs further investigation. One of the reasons could be to revise the base data for prediction, a lot of the administrative level containment measures started in between. For instance, the complete lockdown for three weeks from March 24, 2020, onwards. Similar forecasting is done for India and the hotspots states of India based on actual data from January 30, 2020, till May 15, 2020, and forecast is done for the period of 46 days: 16/05/2020 till 30/06/2020. The model summary for India with the least AIC is presented in Table 7.

Table 7: ARIMA models with all possible values of AIC for India

Confirmed		Deaths		Recovered	
ARIMA model	AIC	ARIMA model	AIC	ARIMA model	AIC
ARIMA(1,2,0)	1421.854	ARIMA(0,2,2)	789.6459	ARIMA(1,2,0)	1282.373
ARIMA(0,2,1)	1426.392	ARIMA(0,2,0)	823.0364	ARIMA(0,2,0)	1313.382
ARIMA(2,2,0)	1423.777	ARIMA(1,2,0)	793.7753	ARIMA(0,2,1)	1293.886
ARIMA(0,2,0)	1441.791	ARIMA(1,2,1)	791.0409	ARIMA(2,2,0)	1283.265
ARIMA(1,2,1)	1423.807	ARIMA(1,2,3)	792.4254	ARIMA(1,2,1)	1284.015

The model which has the least AIC is selected as the best model. The best ARIMA models for forecasting the number of daily confirmed, deaths and recovered cases are ARIMA(1,2,0), ARIMA(0,2,2), ARIMA(1,2,0) respectively for India. Weekly forecasts of COVID-19 with confidence interval are presented in Table 8.

The equation corresponding to the best ARIMA(1,2,0) model for confirmed cases is given by

$$X_t = 19.9474 - 0.4333X_{t-1} + \varepsilon_t \quad (9)$$

The equation corresponding to the best ARIMA(0,2,2) model for death cases is given by

$$Y_t = 0.6201 + \varepsilon_t - 0.6618\varepsilon_{t-1} + 0.2145\varepsilon_{t-2} \tag{10}$$

The equation corresponding to the best ARIMA (1,2,0) model for recovery cases is given by

$$Z_t = 10.904 - 0.5950Z_{t-1} + \varepsilon_t \tag{11}$$

Table 8: Weekly forecast values of COVID-19 with 95% CI for India

Date	Confirmed	Deaths	Recovered
16-05-20	89742 (89335, 90149)	2861 (2841, 2880)	32098 (31888, 32307)
23-05-20	116778 (112499, 121056)	3602 (3432, 3772)	46119 (44118, 48120)
30-05-20	143821 (133520, 154122)	4343 (3939, 4748)	60084 (55293, 64875)
06-06-20	170864 (153005, 188723)	5085 (4387, 5783)	74051 (65762, 82340)
13-06-20	197908 (171220, 224595)	5826 (4785, 6868)	88018 (75644, 100391)
20-06-20	224951 (188327, 261575)	6568 (5140, 7996)	101984 (85016, 118953)
27-06-20	251994 (204438, 299550)	7309.9 (5457, 9162)	115951 (93927, 137975)

Forecasted confirmed COVID-19 cases would be 263584, deaths would be 7627 and recoveries would be 121937 on 30th June 2020. To estimate model adequacy, the Ljung-Box test which is a diagnostic tool is used to test the lack of fit of a time series model. H_0: The model does not show a lack of fit. The alternative hypothesis, H_1: the model does show a lack of fit. Here, for confirmed cases *p*-value is 0.6307, for deaths *p*-value is 0.8192 and for recoveries, *p*-value is 0.1003. As for all the cases, *p*-value is greater than 0.05, hence we do not reject the null hypothesis and conclude that our model does not show a lack of fit.

Now we will forecast the figures for the highly affected states in India assuring that the data is stationary and reliable to forecast. The final models that are reported in table 9 have the lowest AIC values for all hotspot states of India. To estimate model adequacy, the Ljung-Box test which is a diagnostic tool is used to test the lack of fit of a time series model. The output for all six hotspot states of India is given in Table 9. The null hypothesis, H_0: The model does not show a lack of fit. The alternative hypothesis, H_1: The model does show a lack of fit. The *p*-value for state Rajasthan is less than 0.05 for deaths and recoveries, hence we reject the null hypothesis and conclude that model does show lack of fit whereas *p*-value for Rajasthan is more than 0.05 for confirmed cases, hence we do not reject the null hypothesis and conclude that model does not show lack of fit for confirmed cases. For the other hotspot states *p*-value is greater than 0.05 for confirmed, deaths and recovery cases, hence we do not reject the null hypothesis and conclude that our model does not show a lack of fit.

Table 9: The best ARIMA models with least AIC for six hotspots states of India

Hotspot States	Cases	ARIMA Model	AIC	Ljung-Box p-value
Maharashtra	Confirmed	ARIMA(2,2,2)	1338.927	0.7898
	Deaths	ARIMA(2,2,3)	571.9736	0.9937
	Recovered	ARIMA(2,2,2)	1101.577	0.3162
Delhi	Confirmed	ARIMA(2,2,2)	1159.906	0.8447
	Deaths	ARIMA(0,2,2)	449.952	0.9321
	Recovered	ARIMA(1,2,2)	1166.805	0.9672
Madhya Pradesh	Confirmed	ARIMA(1,2,1)	1030.119	0.9634
	Deaths	ARIMA(0,2,1)	456.2507	0.3309
	Recovered	ARIMA(1,2,2)	936.19	0.2378
Tamil Nadu	Confirmed	ARIMA(3,2,0)	1126.77	0.1057
	Deaths	ARIMA(2,2,0)	234.3007	0.8082
	Recovered	ARIMA(2,2,2)	1049.927	0.9991
Gujarat	Confirmed	ARIMA(0,2,1)	1008.264	0.8712
	Deaths	ARIMA(2,2,1)	563.1694	0.9569
	Recovered	ARIMA(4,2,2)	1032.743	0.9571
Rajasthan	Confirmed	ARIMA(1,2,0)	944.7558	0.9514
	Deaths	ARIMA(1,2,2)	396.4093	0.0182
	Recovered	ARIMA(2,2,2)	980.4154	0.0397

Forecast values of ARIMA models with confidence interval for six hotspot states of India are given in table 10.

Table 10: Weekly forecast values of COVID-19 with 95% CI for hotspot states of India

Date	Cases	Maharashtra	Delhi	Madhya Pradesh	Tamil Nadu	Gujarat	Rajasthan
16.05.20	Confirmed	30501 (30233, 30770)	9258 (9144, 9372)	4760 (4697, 4823)	10605 (10507, 10704)	10270 (10213,10327)	4956 (4914,4998)
	Deaths	1116 (1109, 1123)	132 (128, 136)	245 (241, 249)	75 (74, 82)	628 (621, 634)	128 (125, 131)
	Recovered	7064 (6978, 7150)	3874 (3755, 3993)	2407 (2368, 2447)	2832 (2765, 2899)	4119 (4058, 4181)	2810 (2761, 2859)
23.05.20	Confirmed	41154 (39120,43188)	11838 (11152, 12523)	5897 (5401, 6393)	14184 (12888, 15480)	12639 (11977, 13300)	6430 (6013, 6848)
	Deaths	1451 (1380, 1521)	194 (162, 226)	280 (258, 303)	102 (90, 110)	772 (705, 839)	151 (134, 168)
	Recovered	10802 (10045, 11559)	6008 (5158, 6858)	3291 (3054, 3528)	4094 (3712, 4476)	5921 (5314, 6529)	3411 (3098, 3724)
30.05.20	Confirmed	51836 (4669, 56981)	14565 (12730, 16401)	7031 (6029, 8032)	17737 (14685, 20790)	15007 (13412, 16603)	7905 (6903, 8906)
	Deaths	1786 (1609, 1963)	256 (187, 325)	316 (270, 363)	129 (101, 137)	917 (757, 1078)	174 (141, 207)
	Recovered	14514 (12581,16448)	8167 (6333, 10001)	4177 (3554, 4800)	5466 (4463, 6470)	7504 (6200, 8808)	4007 (3277, 4737)
06.06.20	Confirmed	62516 (53417, 71615)	17283 (13927, 20639)	8164 (6572, 9756)	21288 (16043, 26532)	17376 (14609, 20143)	9379 (7644,11113)
	Deaths	2121 (1809, 2433)	319 (205, 433)	352 (276, 428)	157 (108,165)	1062 (784,1340)	198 (145, 250)
	Recovered	18221 (14794, 21648)	1032 (7287,13359)	5063 (3932, 6194)	6846 (5048, 8644)	9192 (7019, 11366)	4604 (3348, 5860)

13.06.20	Confirmed	73196 (59454, 86938)	20002 (14840, 25164)	9298 (7036, 11559)	24838 (17040, 32637)	19745 (15609, 23881)	10853 (8262, 13443)
	Deaths	2456 (1985, 2926)	381 (215, 548)	388 (279, 497)	184 (112, 192)	1207 (792, 1622)	221 (146, 296)
	Recovered	21926 (16745, 27108)	12480 (8056,16903)	5949 (4215, 7684)	8226 (5492, 10959)	10932 (7742, 14122)	5201 (3332, 7070)
20.06.20	Confirmed	2790 (2142, 3439)	22720 (15508, 29932)	10431 (7428, 13434)	28389 (17719, 39059)	22114 (16438, 27790)	12327 (8773, 15881)
	Deaths	2790 (2142, 3439)	444 (219, 668)	424 (278, 570)	212 (113, 219)	1353 (783, 1922)	244 (145, 344)
	Recovered	25631 (18469, 32793)	14636 (8661, 20611)	6835 (4416, 9254)	9605 (5812, 13398)	12625 (8299, 16951)	5798 (3239, 8357)
27.06.20	Confirmed	94557 (69798, 119316)	25438 (15959, 34917)	11564 (7754, 15375)	31939 (18113, 45766)	24482 (17111, 31854)	13801 (9188, 18414)
	Deaths	3125 (2280, 3971)	506 (217, 795)	459 (273, 646)	239 (111, 247)	1498 (759, 2237)	268 (141, 394)
	Recovered	29336 (19990, 38682)	16792 (9119, 24466)	7721 (4545, 10898)	10985 (6023,15947)	14312 (8742, 19882)	6395 (3078, 9712)

7. Discussion and Conclusions

In this paper, we have conducted a two-round study of COVID-19 cases in India and six hotspots states of India. Model accuracy is checked for the first round and then the predication is verified from 15 April 2020 to 25 April 2020. The first-round model is built on data of cumulative confirmed, recovery and death cases from 30 January 2020 to 14 April 2020. We have evaluated the accuracy of the ARIMA model in predicting cumulative confirmed, recovery and death cases. For all six hotspots states of India, the ARIMA model in predicting cumulative confirmed cases is 83% on average which indicates that ARIMA has given good accuracy of prediction. If we discuss country India, forecasted cumulative confirmed cases give 98% model accuracy using the ARIMA model. While model accuracy of cumulative recovery cases and death cases are 97% and 78% respectively. On the other hand, model accuracy for death and recovery cases in six hotspots states of India is 76% and 58% respectively. This seems we need a better model for forecasting death and recovery cases in hotspots states of India. Thus, through this model forecasted confirmed cases are more reliable than with death cases and recovery cases in six hotspots states of India. We hope that our forecasts will be a useful tool for governments and individuals towards making decisions and taking the appropriate actions to curb the spreading of the virus.

There are certain limitations in the numbers of COVID-19 cases forecasted. The forecast is based on past data and information, whereas the technology changes with time and medical science are in the process of doing inventions for the betterment of mankind. If new methods or medicines are invented for the treatment of COVID-19, the figures forecasted may vary. The numbers forecasted may also vary if the effective methods are not adopted or medicines or vaccines are not invented for the treatment of COVID-19 cases. Depending upon the resources, if a greater number of tests are conducted nationwide, the better management of the disease can be done and more spread of disease can be avoided. While considering figures forecasted, we should understand that we have not considered urban-rural variations, stratification of age, occupation, pre-existing co-morbidities, travel history which alters the outcomes. The testing rate is lower in India than in different countries, so our absolute numbers might below. If there is a substantial increase in tests, it may also affect the numbers forecasted. If healthcare facilities are increased, the forecasted figures may alter.

Acknowledgements

The authors express their gratefulness to the reviewer and the editor for their invaluable comments and suggestions, which have helped us to improve the paper substantially.

References

Kai, L., Ying, C., Ruzheng, L. and Kunyuan, H. (March, 2020). Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients. *Journal of Infection*, **6,** *e*14-*e*18.

Khot, W. Y. and Nadkar, M. Y. (March, 2020). The 2019 novel coronavirus outbreak – A global threat. *Journal of the Association of Physicians of India*, **68**.

Khrapov, P.V. and Loginova, A.A. (2020). Mathematical modelling of the dynamics of the Coronavirus COVID-19 epidemic development in China. *International Journal of Open Information Technologies*, **8(4),** 13-16.

Lin, J., Kewen, L., Jiang, Y., Xin, G. and Ting, Z. (March, 2020). Prediction and analysis of coronavirus disease 2019. *arXiv.org>q-bio>arXiv*, 2003.05447.

Lina, Q., Zhaob, S., Daozhou, G., Loue Y., Shu, Y., Musa, S., Wangb, M. H., Caig, Y., Wang, W., Yangh, L. and Hee, D. (March, 2020). A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *International Journal of Infectious Diseases*, **93**, 211-216.

Liu, Y., Gayle A. A., Wilder-Smith, A. and Rocklöv, J. (February, 2020). The reproductive number of COVID-19 is higher compared to SARS Coronavirus. *Journal of Travel Medicine*, **27,** 1-4.

Miller, A., Reandelar, M. J., Fasciglione, K., Roumenova, V., Li, Y. and Otazu, G. H. (March 2020). Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study. *CC-BY-ND 4.0 International license.* https://doi.org/10.1101/2020.03.24.20042937.

Petropoulos, F. and Makridakis, S. (March, 2020). Forecasting on the novel coronavirus COVID-19. *PLOS ONE*, https://doi.org/10.1371/journal.pone.0231236.

Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A. and Jewell, C. P. (January, 2020). Early prediction of the 2019 novel coronavirus outbreak in the mainland China based on simple mathematical model. *IEEE*, **8,** 51761-51769.

Ribeiro, M.H.D.M., Da Silva, R.G., Mariani, V.C. and Coelho, L.S. (May, 2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons and Fractals,* **135,** 1-10.

Schueller, E., Klein, E., Tseng, G. K., Balasubramanian, R., Kapoor, G., Joshi, J., Sriram, A., Nandi, A. and Laxminarayan, R. (April, 2020). COVID-19 in India: Potential Impact of the Lockdown and Other Longer-Term Policies. *The Centre for Disease Dynamics, Economics and Policy.*

Tania D., Mardani-Fard, H.A. and Paria, D. (March 2020). Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA. *MedRxiv,* preprint. doi: https://doi.org/10.1101/2020.03.13.20035345.