



# Time Series Forecasting for Coronavirus (COVID-19)

Priyal Sobti<sup>1</sup>, Anand Nayyar<sup>2</sup>(✉), and Preeti Nagrath<sup>1</sup>

<sup>1</sup> Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi 110063, India

<sup>2</sup> Graduate School, Faculty of Information Technology, Duy Tan University, Da Nang 550000, Vietnam  
anandnayyar@duytan.edu.vn

**Abstract.** The upsurge of the novel coronavirus has spread to many countries and has been declared a pandemic by WHO. It has shaken the most powerful countries across the world like the USA, UK, and has affected economies of various countries. The coronavirus or the 2019-nCoV causes the disease that has been named COVID-19. This disease transmits by inhaling droplets that are expelled by an infected person. It has been affecting people in different ways and has been found to be threatening for the older population or people with comorbidities. It has been seen that the virus 2019-nCoV spreads faster than the two of its antecedents namely severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV). No cure or vaccine has been discovered as of now and taking precautions like staying at home are the only possible solutions.

Our study analyzes the current trend of the disease in India and predicts future trends using time series forecasting. The official dataset provided by John Hopkins University through a GitHub repository has been used for the research for the time period of 22 January 2020 to 31 May 2020. The trend in cases, fatalities, and the people who have recovered until the date of 31 May 2020 has been discussed in the paper. It has been seen through the findings that the total number of cases is expected to rise to 2,15,000 by the end of May 2020 i.e. 31 May 2020 as per the AR (Autoregression) model. ARIMA (Autoregressive Integrated Moving Average) model predicts the number of cases to be 2,05,000 until the same date. Actual data has shown that the number of confirmed cases is 1,90,609 as on 31 May 2020 giving a percentage error of 7.57% and 12.85% for ARIMA and AR model respectively. Comparison between the findings of the two models has been shown later in the paper.

**Keywords:** COVID-19 · Coronavirus · SARS-CoV-2 · Lockdown · Social distancing · Time series forecasting · Time series analysis

## 1 Introduction

COVID-19 is an infectious disease caused by a novel virus, SARS-CoV-2 or 2019-nCoV. It has symptoms like dry cough, fever, breathing difficulties (Singhal 2020) and

in severe cases it has been found to be fatal for older people and people with pre-existing medical conditions. The only technique believed to slow the spread of the virus is social distancing and hence, the extended lockdown in India till 17 May 2020. It has infected over 2.2 million people worldwide as of 18 April 2020 and over 1,50,000 fatalities (W.H.O. 2020) have been reported till the same date. It is a highly contagious disease and the only way to protect ourselves is by social distancing and taking care of our personal hygiene like washing our hands for at least 20 s. The virus allows for spread to take place very quickly as it has an incubation period ranging from 7 days to 14 days and in some cases, it has even reached 24 days. The first case of the disease had been reported in India on 30 January 2020. Since then, the number of infections has been on the rise and has recently begun to slow down in some states. The worst affected states remain Maharashtra, Delhi, and Tamil Nadu.

Across the world, the worst affected countries are namely Italy, France, Spain, Britain, the United States of America with New York as the epicenter for the outbreak. The disease has taken a toll on the health care systems of the countries and it goes to show that no country was prepared for the pandemic. The massive fatalities tolls in developed countries are indicative of the overburdened healthcare systems and the fact that the services are not able to reach the people who need them the most.

In our study, we aim to analyze the trend and make predictions for the number of cases, fatalities and recoveries in India using time series forecasting. Time series is a collection of information that is accumulated sequentially over time (Chatfield 2000). Time series forecasting makes predictions using data made up of time series. Time series forecasting makes forecasts using past and present values. Time series forecasting is regularly used in the field of sales forecasting (Ansuji et al. 1996), inventory control, capacity planning, budgeting, financial markets (Tay and Cao 2001) and much more (Chatfield 2000). Time series model provides predictions for the future and this gives us an opportunity to compare the results with the actual observed values. Models namely ARIMA (Benvenuto et al. 2020) (Dehesh et al. 2020) and AR have been used to carry out time series forecasting and make predictions for the number of cases till the end of May 2020. ARIMA techniques have been regularly used due to their accuracy for forecasting as well as time series analysis (Contreras et al. 2014). ARIMA model is a class of models that are used to study the past values of a time series and predict future values. Also, ARIMA is a combination of two processes i.e. an Auto Regressive (AR) process and a Moving Average process and models based on these processes are also available. One of the models namely AR model has also been used in this study.

The objectives of the paper are:

- To use time-series analysis to analyze the trend using the data provided by John Hopkins University (John Hopkins University Dataset 2020 (accessed on May 12, 2020)).
- To use time-series forecasting to predict the trends in cases, fatalities and recoveries using different time series forecasting models like ARIMA.
- To compare the trends and the predictions made by two time series forecasting models namely ARIMA and AR.
- To analyze the results obtained from the two models to understand the situation and draw conclusions from it.

The paper consists of 5 sections. Section 2 highlights the literature review conducted in the domain of research. Section 3 focuses on the methodology used to solve the problem at hand. Section 4 highlights and showcases the results obtained and the analysis of the results. Section 5 contains the conclusion for the paper.

## 2 Literature Review

The researchers and academicians across the world are carrying out extensive research in this field and some of them are as follows. Arti et al. (2020) have suggested a tree-based model in which some people are quarantined and some are left undetected due to lack of symptoms, hiding travel history, etc. They have shown the effect of lockdown by considering different scenarios and the number of days.

Gupta et al. (2020) have performed exploratory data analysis and have used time-series forecasting to predict future trends. According to one such model used; they have shown that 3 million people may get infected if proper measures are not taken. They had conducted their study when the number of cases in India was 536 and were expected to rise to 7000 as per the ARIMA time series forecasting method.

Baud et al. (2020) have re-estimated the mortality rates. They have obtained the mortality rates by dividing the number of deaths on a given day by the number of patients that have contracted the disease 14 days prior. Their findings have suggested a mortality rate of 5.6% in China and 15.2% outside China.

Deb et al. (2020) went onto show the effect of partial and total lockdown by proposing a time series model.

Petropoulos et al. (2020) showed using forecasting that the number of cases is expected to rise. They have shown the trajectories for the reported cases as well as the recovered cases and have analyzed the same for different time periods since the outbreak.

Healthcare impact due to COVID-19 has been shown by Chatterjee et al. (2020) wherein they have developed a compartmental SEIR model. Aspects like patient hospitalization, the requirement of Intensive Care Units (ICUs) had been modelled using SimVoi software. They have concluded by suggesting that the Indian healthcare might be overwhelmed by the end of May.

Pandey et al. (2004) used two models namely SEIR and Regression models to predict the trend and the changes in trend in the COVID-19 spread. They found that the number of cases till 30 March 2020 will be less than 0.5 per million in India.

Age-related impact of social distancing on the COVID-19 pandemic in India has been discussed by Singh et al. (2020). They went onto suggest a mathematical model of the disease transmission that takes into account both the age and the social contact structure. They have emphasized the fact that while keeping a check on the total number of cases, it is also required to take into account the age group affected. The paper also showed that how the three-week lockdown in India starting from 25 March 2020 is insufficient and also suggested some ways to contain the spread of the virus including extending the lockdown with periodic relaxations.

Roy et al. (2020) have conducted a study to understand the notions and the thoughts of Indian population when it comes to coronavirus. The study shows that people are appropriately aware of preventive measures. However, people have apprehensions regarding

their mental health during this time and are predominantly worried about catching the infection.

Drug repurposing is used regularly nowadays where drugs that are originally developed for some other diseases are used to treat other diseases. Muralidharan et al. (2020) have tried to understand the mechanism of the proposed drugs namely lopinavir, oseltamivir and ritonavir which are being used to reduce the virulence in the infected patients.

Sahoo et al. (2020) have also highlighted the mental distress that is being caused by various factors like fear of infection, economic loss, unemployment, lack of social interaction and much more. They have presented two such case studies which had been reported to their medical services.

Tanne et al. (2020) have discussed how doctors and healthcare systems are dealing with the pandemic. They have discussed the impact in various countries like India, the U.S.A, Spain, Japan, South Korea and many more in terms of the number of cases and the subsequent response of the country to tackle the pandemic.

LSTM techniques have been proposed by Tomar et al. (2020) and estimation for the next 30 days for the number of positive cases in India has been provided. Also, the effect of measures like lockdown and social distancing has been discussed.

The use of epidemiological models has been shown by Rajesh Ranjan (2020) for the prediction for COVID-19 outbreak. Susceptible-infected recovered (SIR) and exponential models were used to make predictions using the known data and it has been seen that India will enter equilibrium by the end of May provided there is no community transmission.

Vellingiri et al. (2020) have discussed the symptoms for COVID-19 and how the symptoms are different from SARS, MERS and common flu. Also, drugs available, treatment methodology and ongoing vaccine trials have also been discussed. Some traditional Indian medicinal plants that can provide some therapeutic relief have also been discussed.

Time series analysis and forecasting have been used as the main technique for predicting confirmed cases for COVID-19. Time Series forecasting as defined by Chatfield et al. (2000) is predicting values on the basis of data constituting time series. Gooijer et al. (2006) have assessed the past 25 years of research that have gone into time series forecasting and have reviewed some influential work in this area.

### 3 Methodology

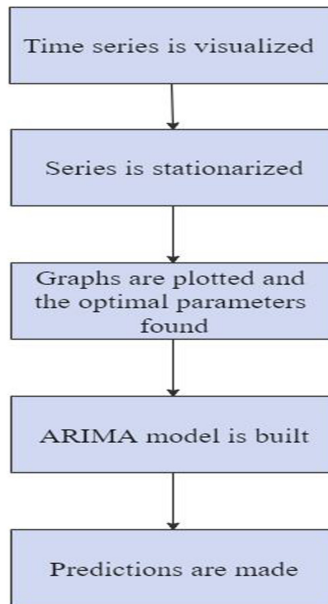
Time series analysis is the technique to analyze the time-series data to find characteristics and statistics from the data (Chatfield 2000). On the other hand, time series forecasting (Chatfield 2000) is the technique to predict future values from the existing data values.

The first step followed for time series forecasting included data pre-processing. The data has been taken from the official GitHub repository provided by John Hopkins University which is daily updated with the latest numbers. Since the study is focused on cases in India, the relevant row was chosen and the data frame obtained after that was transposed as the dates had to be taken as a column. The data is then converted into a Time Series object with the dates as the index. We can also visualize the time

series data using time series decomposition that allows decomposition into three distinct components namely trend, seasonality, and noise.

The model used for time series forecasting includes Autoregression (AR) and Autoregressive Integrated Moving Average (ARIMA). ARIMA is the most commonly used model for time series forecasting and includes parameter selection that shall yield the best possible outcome. After choosing the optimal combination of parameters the model is fit and the corresponding plots are obtained. After analyzing the given trend and preparing the model, appropriate functions like `predict()` were used to predict the future behavior of the cases, fatalities, and recoveries from 13 May 2020 to 31 May 2020.

Figure 1 represents the steps that are followed in the methodology.



**Fig. 1.** Flowchart for methodology

## 4 Experimental Analysis

### 4.1 Dataset

The dataset provided by John Hopkins University (John Hopkins University Dataset 2020 (accessed on May 12, 2020)) in the form of a GitHub repository has been used. The repository is regularly updated with the latest numbers. The dataset contains the number of confirmed cases, fatalities as well as the number of recoveries for countries globally. This study makes use of numbers for India and the trend in confirmed cases, fatalities and recoveries is visualized.

### 4.2 Plot Analysis

The time-series data created was first plotted to see the general trend for the number of cases, recoveries, and fatalities. Figure 2 shows the current trend of cases in India. Similar plots can be seen for the number of deaths and recoveries in India through Figs. 3 and 4 respectively.

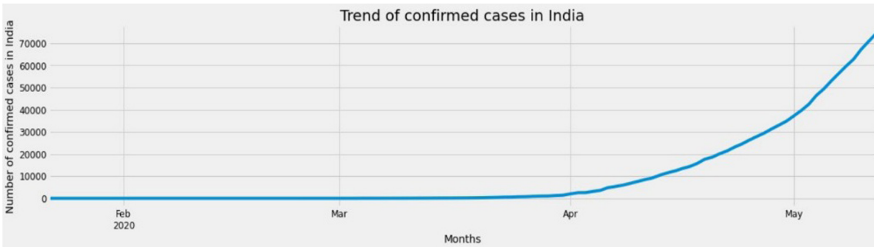


Fig. 2. The trend of confirmed cases in India

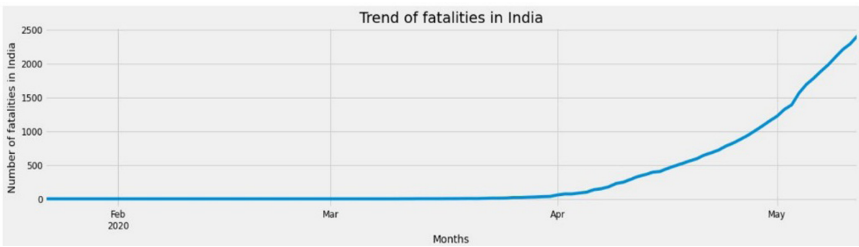


Fig. 3. The trend of the number of fatalities in India

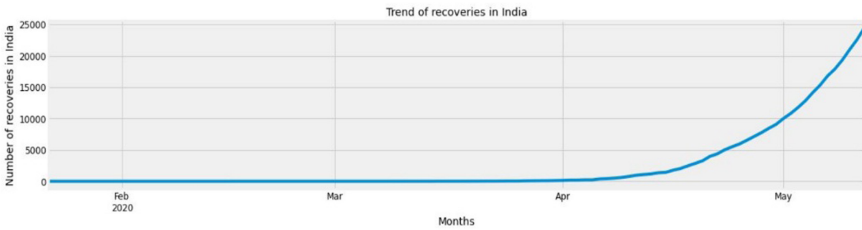
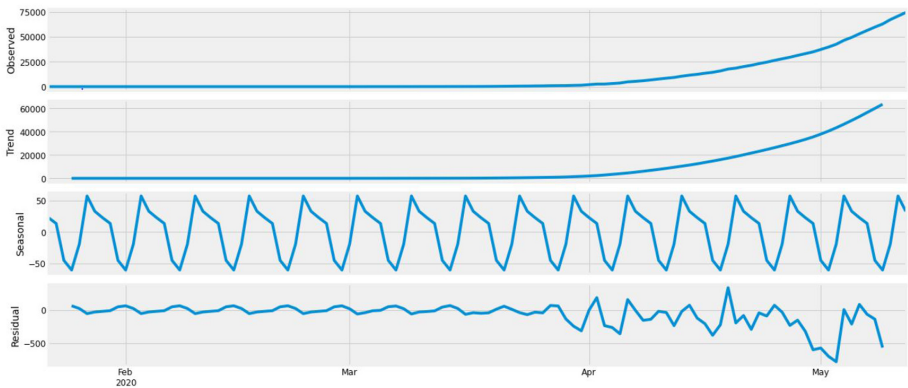


Fig. 4. The trend of the number of recoveries in India

Time series data can be decomposed into three components namely trend, seasonality, and residual with the help of a built-in library function. Any time series can be broken into systematic or non-systematic components where systematic components are the ones that have some form of recurrence while non-systematic components are otherwise. Level, trend, and seasonality are systematic components while noise is a non-systematic component. These components can be either added together called the additive model or they can be multiplied together called the multiplicative model.

Seasonal decompose () is an inbuilt function that allows for automatic time series decomposition by plotting the original data, trend, seasonality, and residual components which is nothing but the time series after the trend and the seasonal component is removed.

Figure 5 shows the decomposition for the time series depicting the number of cases in India. The seasonal component shows the short-term cycle that repeats itself in terms of the number of cases. The residual component shows how the cases had been very less during the initial days of the month but around the end of March and the beginning of April, there had been a spike in the number of cases.



**Fig. 5.** Decomposition of the time series depicting the number of cases in India

### 4.3 Results and Discussion

Using the Autoregression (AR) model the future trend has been predicted from 13 May 2020 to 31 May 2020. The plot still shows an upward trend in the number of cases as well as the number of recoveries and fatalities. All these predictions have been made in the mathematical sense and social elements like the fear of the disease, people adhering to lockdown rules have not been taken into consideration. Figure 6 shows the prediction for the number of cases that are still expected to rise. The prediction shows that the number of cases will be as high as 2,15,000 by the end of May 2020 considering the current trend in the number of cases. Figure 7 and Fig. 8 shows the forecast for the number of fatalities and recoveries as per the AR model.

Similarly, the ARIMA model has also been used to do similar predictions as it is considered to be one of the most commonly used models for time series forecasting. Figure 9 shows the plot for the number of confirmed cases as per the ARIMA model. Also, Fig. 10 and Fig. 11 shows the trend for the number of fatalities and the number of recoveries as per the ARIMA model for COVID-19. All the figures that have been obtained for the two models show an exponential increase in the number of confirmed cases as well as the number of recoveries and fatalities for the increasing x-axis which is time.

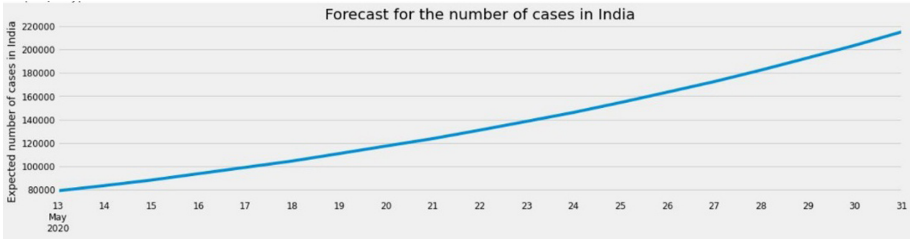


Fig. 6. Forecast for confirmed cases as per the AR model

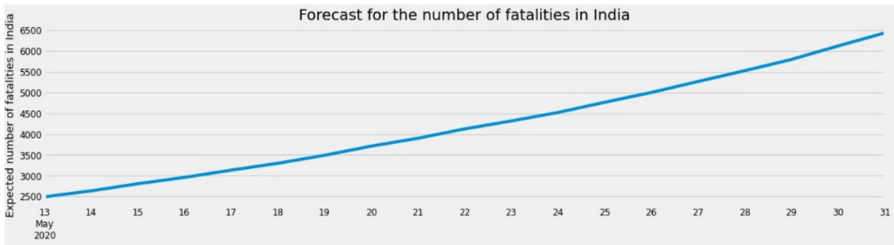


Fig. 7. Forecast for number of fatalities as per the AR model

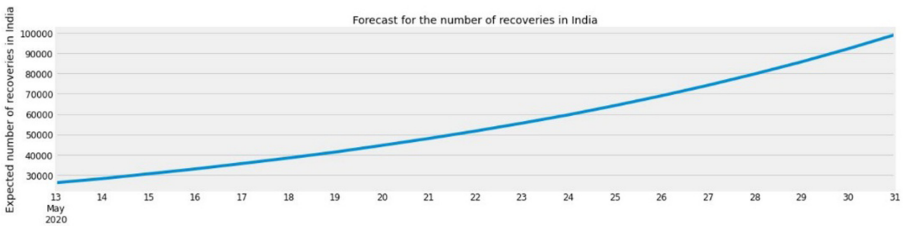


Fig. 8. Forecast for number of recoveries as per the AR model

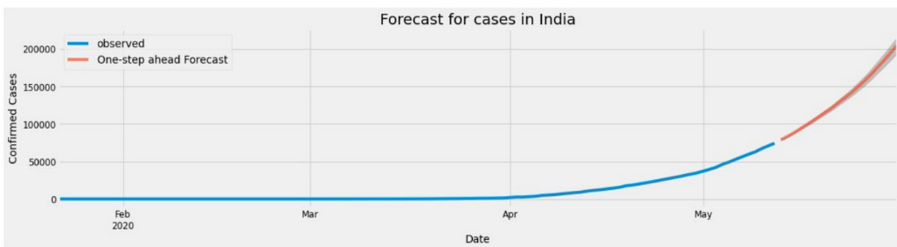
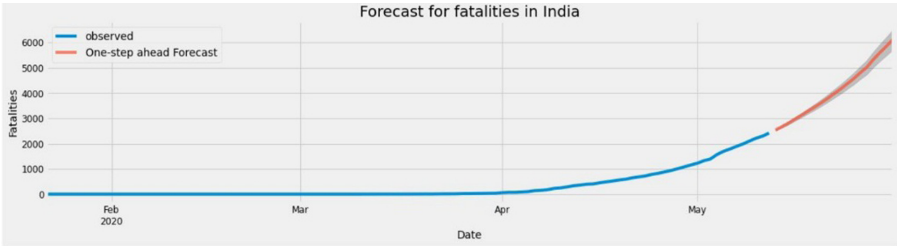
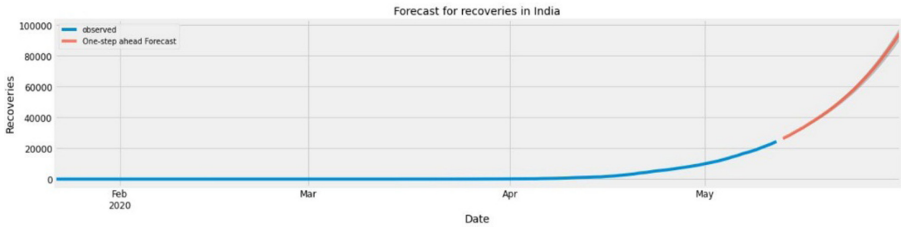


Fig. 9. Forecast for confirmed cases as per the ARIMA model





**Fig. 10.** Forecast for number of fatalities as per the ARIMA model



**Fig. 11.** Forecast for number of recoveries as per the ARIMA model

ARIMA model is a combination of two processes namely an Auto Regressive (AR) process and a Moving Average process and models based on these processes are also available. Equation 1 represents the equation for predicting values using the AR model:

$$Y = \beta_0 + \beta_1 * l_1 + \beta_2 * l_2 + \dots + \beta_n * l_n \tag{1}$$

where Y represents the observed value,  $\beta_1, \beta_2 \dots$  are regression betas and  $l_1, l_2, \dots$  represent the lag values. A similar equation can be given for the MA model. Equation 2 represents the equation for predicting values using the MA model:

$$Y = \beta_0 + \beta_1 * e_{-l_1} + \beta_2 * e_{-l_2} + \dots + \beta_n * e_{-l_n} \tag{2}$$

where  $e_{-l_1}, e_{-l_2}$  and so on represent random residual deviations between the MA model and target variable.

ARIMA (p, d, q) is a non-seasonal ARIMA model where p is number of autoregressive terms, q is the number of forecast errors in the prediction equation and d is the number of nonseasonal differences. Equation 3 represents the general forecasting equation used by the ARIMA model to predict future values.

$$y_t = \mu + \alpha_1 y_{t-1} + \dots + \alpha_p y_{t-p} - \theta_1 e_{t-1} - \dots - \theta_q e_{t-q} \tag{3}$$

where  $y_t$  represents the data on which ARIMA has to be applied,  $\alpha_1, \alpha_2 \dots$  are AR coefficients and  $\theta_1, \theta_2$  and so on are MA coefficients.

Table 1 summarizes and compares the predictions that have been made by the two models from 13 May 2020 to 31 May 2020. It can be seen that there is a rise in the total number of cases and both the models go to show that the total number of cases will more than 2,00,000 by the end of May.

**Table 1.** Prediction of the number of COVID-19 patients in India by the end of May.

Date	AR Model	ARIMA Model
2020-05-13	78989	78944
2020-05-14	83470	83269
2020-05-15	88202	88193
2020-05-16	93646	93704
2020-05-17	99064	99034
2020-05-18	104479	104698
2020-05-19	110845	110525
2020-05-20	117303	116414
2020-05-21	123672	122308
2020-05-22	130924	129153
2020-05-23	138428	135706
2020-05-24	146061	142415
2020-05-25	154542	150172
2020-05-26	163414	157964
2020-05-27	172424	166435
2020-05-28	182315	175950
2020-05-29	192787	185117
2020-05-30	203471	194975
2020-05-31	215028	205039

## 5 Conclusion and Future Scope

In this study, two time series forecasting models namely AR and ARIMA models were used to analyze and predict the number of confirmed cases for COVID-19 in India. The data provided by John Hopkins University included the data for the COVID-19 pandemic worldwide. Therefore, proper techniques had been applied to extract the data for India and carry out time series forecasting till the end of May 2020. The data has been analyzed and it has been found out that the number of cases is expected to rise till the end of May 2020. Both the ARIMA and AR model point towards an increasing number of cases as well as an increasing number of fatalities and recoveries. Table 1 shows that the number of confirmed cases in India will cross 2,00,000 by the end of May 2020. The results from this study have been concluded using the data until the 12 May 2020. Actual data has shown that the number of confirmed cases in India are 1,90,609 as on 31 May 2020 giving a percentage error of 7.57% for the ARIMA model and 12.8% for the AR model. The trend shows an exponential upsurge in the number of cases and this can aggravate due to negligence of individuals or community spread. Furnishing the hospital with proper medical equipment can help tackle the pandemic efficiently.

In the future, other time series forecasting models can be used to accurately predict the trend by taking into consideration other factors like adherence to lockdown rules, people following the required precautions etc. as the results obtained can help medical authorities better prepare for the situation.

## References

- Ansuji, A.P., Camargo, M., Radharamanan, R., Petry, D.: Sales forecasting using time series and neural networks. *Comput. Ind. Eng.* **31**(1–2), 421–424 (1996)
- Arti, M., Bhatnagar, K.: Modeling and predictions for covid 19 spread in India. *ResearchGate* (2020)
- Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L., Favre, G.: Real estimates of mortality following COVID-19 infection. *The Lancet Infectious Diseases* (2020)
- Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., Ciccozzi, M.: Application of the arima model on the Covid-2019 epidemic dataset. *Data Brief*, 105340 (2020)
- Chatfield, C.: *Time-Series Forecasting*. CRC Press, Boca Raton (2000)
- Chatterjee, K., Chatterjee, K., Kumar, A., Shankar, S.: Healthcare impact of Covid-19 epidemic in India: a stochastic mathematical model. *Med. J. Armed Forces India* (2020)
- Contreras, J., Espinola, R., Nogales, F.J., Conejo, A.J.: Arima models to predict next-day electricity prices. *IEEE Trans. Power Syst.* **18**(3), 1014–1020 (2003)
- Deb, S., Majumdar, M.: A time series method to analyze incidence pattern and estimate reproduction number of covid-19. *arXiv preprint arXiv:2003.10655* (2020)
- De Gooijer, J.G., Hyndman, R.J.: 25 years of time series forecasting. *Int. J. Forecast.* **22**(3), 443–473 (2006)
- Dehesh, T., Mardani-Fard, H., Dehesh, P.: Forecasting of covid-19 confirmed cases in different countries with arima models. *medRxiv* (2020)
- Gupta, R., Pal, S.K.: Trend analysis and forecasting of covid19 outbreak in India. *medRxiv* (2020) John Hopkins university dataset. (2020). <https://github.com/CSSEGISandData/COVID-19>. Accessed 12 May 2020
- Muralidharan, N., Sakthivel, R., Velmurugan, D., Gromiha, M.M.: Computational studies of drug repurposing and synergism of lopinavir, oseltamivir and ritonavir binding with sars-cov-2 protease against covid-19. *J. Biomolecular Struct. Dyn.* **56** 1–6 (2020)
- Pandey, G., Chaudhary, P., Gupta, R., Pal, S.: Seir and regression model based covid-19 outbreak predictions in India. *arXiv preprint arXiv:2004.00958* (2020)
- Petropoulos, F., Makridakis, S.: Forecasting the novel coronavirus covid-19. *PLoS ONE* **15**(3), (2020)
- Ranjan, R.: Predictions for covid-19 outbreak in India using epidemiological models. *medRxiv* (2020)
- Roy, D., Tripathy, S., Kar, S.K., Sharma, N., Verma, S.K., Kaushal, V.: Study of knowledge, attitude, anxiety & perceived mental healthcare need in Indian population during covid-19 pandemic. *Asian J. Psychiatry*, 102083 (2020)
- Sahoo, S., et al.: Self-harm and covid-19 pandemic: an emerging concern—a report of 2 cases from India. *Asian J. Psychiatry* (2020)
- Singh, R., Adhikari, R.: Age-structured impact of social distancing on the covid-19 epidemic in India. *arXiv preprint arXiv:2003.12055* (2020)
- Singhal, T.: A review of coronavirus disease-2019 (covid-19). *The Indian J. Pediatrics*, 1–6 (2020)
- Tanne, J.H., Hayasaki, E., Zastrow, M., Pulla, P., Smith, P., Rada, A.G.: Covid-19: how doctors and healthcare systems are tackling coronavirus worldwide. *BMJ*, 368 (2020)

- Tay, F.E., Cao, L.: Application of support vector machines in financial time series forecasting. *Omega*, **29**(4), 309–317 (2001)
- Tomar, A., Gupta, N.: Prediction for the spread of covid-19 in India and effectiveness of preventive measures. *Sci. Total Environ.* 138762 (2020)
- Vellingiri, B., et al.: Covid-19: a promising cure for the global panic. *Sci. Total Environ.* 138277 (2020)
- W.H.O.: Coronavirus disease 2019 (Covid19): situation report (2020)