Modeling Earth Systems and Environment (2021) 7:29–40
https://doi.org/10.1007/s40808-020-01080-6

ORIGINAL ARTICLE

Mathematical modeling and a month ahead forecast of the coronavirus disease 2019 (COVID-19) pandemic: an Indian scenario

Suhail Ganiny[1] · Owais Nisar[2]

Received: 4 September 2020 / Accepted: 31 December 2020 / Published online: 19 January 2021

Abstract

India, the second-most populous country in the world is witnessing a daily surge in the COVID-19 infected cases. India is currently among the worst-hit nations worldwide due to the COVID-19 pandemic and ranks just behind Brazil and the USA. The prediction of the future course of the pandemic is thus of utmost importance in order to prevent further worsening of the situation. In this paper, we develop models for the past trajectory (March 01, 2020–July 25, 2020) and also make a month-long (July 26, 2020–August 24, 2020) forecast of the future evolution of the COVID-19 pandemic in India by using an autoregressive integrated moving average (ARIMA) model. We determine the most optimal ARIMA model (ARIMA(7,2,2)) based on the statistical parameters viz. root-mean-squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2). Subsequently, the developed model is used to obtain a one month-long forecast for the cumulative cases, active cases, recoveries, and the number of fatalities. According to our forecasting results, India is likely to have 3800,989 cumulative infected cases, 1634,142 cumulative active cases, 2110,697 cumulative recoveries, and 56,150 cumulative deaths by August 24, 2020, if the current trend of the pandemic continues to prevail. The implications of these forecasts are that in the upcoming month, the infection rate of COVID-19 in India is going to escalate, while the rate of recovery and the case-fatality rate is likely to reduce. In order to avert these possible scenarios, the administration and health-care personnel need to formulate and implement robust control measures, while the general public needs to be more responsible and strictly adhere to the established and newly formulated guidelines in order to slow down the spread of the pandemic and prevent it from transforming into a catastrophe.

Keywords COVID-19 · Modeling · Forecast · India · ARIMA

Introduction

Since the first emergence of the Coronavirus Disease 2019 (COVID-19) in Wuhan, Hubei Province, China, in December 2019 (Wu et al. 2020; Zhou et al. 2020; Zhu et al. 2020), the disease has proliferated globally and has affected 215 countries till date (https://www.worldometers.info/coronavirus/). The causative agent of the disease has been identified to be novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) that shares a 79.6% sequence match with its predecessor SARS-CoV (Zhu et al. 2020). SARS-CoV-2 is very likely to have a zoonotic origin, possibly from bats, as it has a similarity of 97% with SARS-like bat CoVs at the whole-genome level (Zhou et al. 2020; Wu et al. 2020). It is however, quite probable that pangolins acted as the intermediate hosts prior to human transmission (Zhang et al. 2020). SARS-CoV-2 is the seventh pathogen belonging to the class of coronaviruses that tend to affect humans, the other six being HCoV-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU1, SARS-CoV and MERS-CoV (Su et al. 2016).

The human-to-human transmission of COVID-19 predominantly occurs through respiratory droplets (5

✉ Suhail Ganiny
suhail_15phd16@nitsri.ac.in

Owais Nisar
owais703777@gmail.com

[1] Mechanical Engineering Department, National Institute of Technology Srinagar, Hazratbal, Srinagar, J&K 190006, India

[2] College of Agricultural Engineering and Technology, Sher-e-Kashmir University of Agricultural Science and Technology, Shalimar, Srinagar, J&K 190025, India

μm < size < 10 μm) or aerosols (size ≤ 5μm), close interpersonal physical contact or by touching infected surfaces (https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations). Very recently, Morawska and Cao (2020) have also acknowledged air borne transmission of COVID-19. Once a person is infected, it is typically characterized by symptoms like fever, cough, fatigue, myalgia, chest pain, dyspnoea and sore throat (Huang et al. 2020). The global outbreak and the severity of this contagious disease (transmissibility rate—$R_0 = 1.4–3.9$ Li et al. 2020) prompted the World Health Organization (WHO) to declare it as a Public Health Emergency of International Concern (PHEIC) on January 30, 2020, and then subsequently, to classify it as a pandemic on March 11, 2020 (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen). Moreover, the COVID-19 R_0 is estimated to be as high as five for events like weddings, religious gatherings, conferences and in industrial settings (Saidan et al. 2020), and thereby such events tend to accelerate the propagation of the disease. Owing to non-availability of a specific vaccine, the management of the disease requires the adoption of measures like social distancing, frequent hand washing, sanitizing, wearing face masks, extensive screening and testing, contact tracing, isolation and quarantining (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public).

According to the current statistics (July 28, 2020, 18:48 GMT), COVID-19 has hitherto affected more than 16 million people (16,779,951) globally, including greater than 10 million survivors (10,333,138), while more than 0.6 million people (660,318) have unfortunately succumbed to the disease (https://www.worldometers.info/coronavirus/). The recovery rate and the case-fatality rate currently stand at 61.58% and 3.93%, respectively. The world is presently witnessing a daily surge of nearly 0.2 million newly infected cases and about 5000 fatalities. The USA, Brazil, India, Russia and South Africa are being the five most worst affected nations at this moment with a case share of 27%, 15%, 9%, 5% and 3% of the total global cases, respectively, whereas the proportion of deaths in these countries is 23%, 14%, 6%, 2% and 1% of the total worldwide deaths, respectively. An overview of the relative spread of the pandemic among the ten heavily affected nations in the world is shown in Fig. 1. The spread and fatality of this continuing pandemic differ stochastically from country to country as a multitude of factors like government response, economic status, testing rate, healthcare infrastructure, environmental conditions, demographics, faithful reporting, compliance with advised measures, etc., contributes to it (Sarmadi et al 2020; Omori et al. 2020). The rapid outbreak of the pandemic has disrupted the normal life of the human inhabitants as we are being strictly adhering and adapting to measures like indoor confinement, nationwide lockdowns, social distancing, travel restrictions, administrative surveillance, limited outdoor activity owing to the closure of workplaces, educational institutes, restaurants, parks, gymnasiums, etc. (de Haas et al. 2020; Di Renzo et al. 2020).

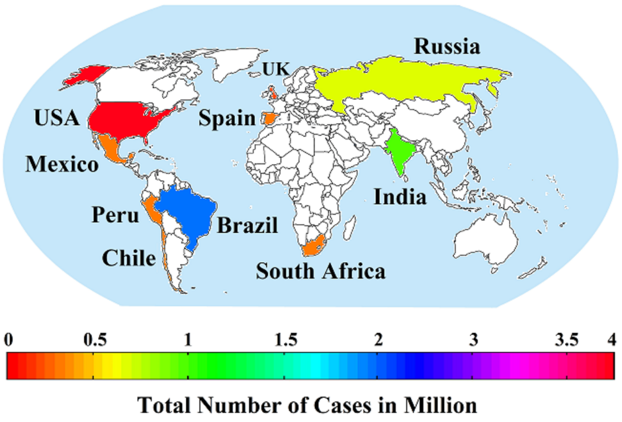

Fig. 1 Relative spread of the COVID-19 pandemic in the top 10 worst affected countries as on July 28, 2020

The first case of COVID-19 in India was reported on January 30, 2020, in Kerala (a coastal state in the south-western part of India), when a student returnee from Wuhan, China, tested positive for the virus (https://www.cnbc.com/2020/01/30/india-confirms-first-case-of-the-coronavirus.html). The second and the third cases were reported from the same state on February 2, 2020 and February 3, 2020, respectively, with both the patients having a Chinese travel history. The ongoing outbreak, however, started in March 2020, when the indigenous cases started surfacing. India has ever since seen a continuing daily increase in the newly identified cases, with the current daily count being nearly 50,000. Although initially the spread of the disease was relatively slow, but gradually the transmission started picking up pace and as of now (July 28, 2020, 18:48 GMT), the total number of diagnosed cases has surged past 1.5 million (1532,125). The diagnosed cases include about 1 million recoveries (988,768), more than 0.5 million active cases (509,133) and 34,224 deaths (https://www.worldometers.info/coronavirus/, https://www.mohfw.gov.in/). With such high numbers, India is currently the worst affected Asian country and the third worst hit country worldwide by the pandemic, and it accounts for nearly 38% of the identified cases in Asia and about 9% of the global cases. India currently has the highest infection rate globally as is evident by the fact that the cumulative infected cases have increased nearly by 20% since the last week and by 65% since the past one month only.

India's journey from the first inception of the virus till now is shown in Fig. 2. India is emerging as the latest global hotspot of the pandemic owing to the high rise in the cases and the fact that most of the population of the

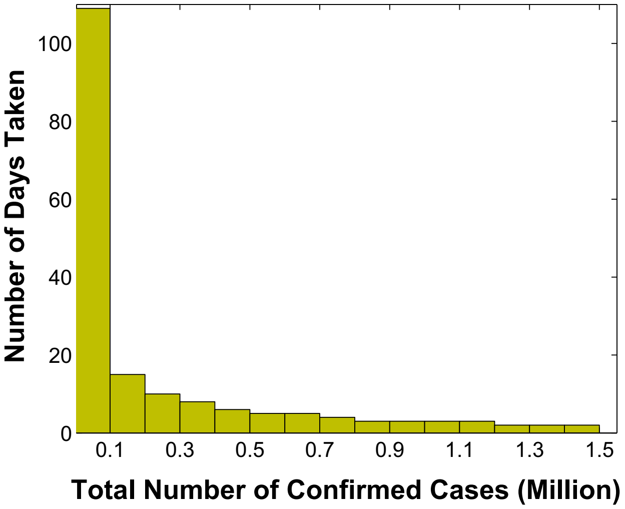

Fig. 2 India's Journey to 1.5 million cases—India took 109 days to reach 0.1 million, followed by 15 days to 0.2 million and just 2 days to reach from 1.4 million to 1.5 million

country lives in densely packed cities. The recovery rate of 64.53% and the case-fatality rate of 2.23% in India are, however, promising and are better than the global values of 61.58% and 3.93%, respectively. It is pertinent to mention that the actual number of infected cases and the mystical low number of fatalities could in reality be much higher as India's testing rates are one of the lowest (12,848 per million population https://www.worldometers.info/coronavirus/). Besides this, the relatively poor health-care infrastructure of the country, and the fact that a large proportion of the population dies in rural areas without any significant medical contemplation, renders their determination and reporting less likely.

COVID-19 has swept across all the 28 states and 8 union territories of India. The infected cases have particularly escalated in the sates of Maharashtra (391,440), Tamil Nadu (227,688), Andhra Pradesh (110,297), Karnataka (107,001), Uttar Pradesh (73,951), West Bengal (62,964), Gujarat (57,982), Telangana (57,142) and Bihar (43,591) (https://www.mohfw.gov.in/). These ten worst affected states account for the almost 74% of the total diagnosed cases in India. The Indian Government has not yet declared the community transmission of COVID-19 in the country; however, the states of Assam, Kerala and West Bengal have announced the same (https://theprint.in/health/india-not-in-community-transmission-stage-as-only-49-districts-account-for-80-cases-govt/457771/). In contadiction to this, many experts are of the opinion that the country is indeed in the community transmission phase of the pandemic as the sources of many diagnosed patients cannot be traced (https://scroll.in/latest/967936/coronavirus-situation-in-india-really-bad-community-transmission-taking-place-says-ima).

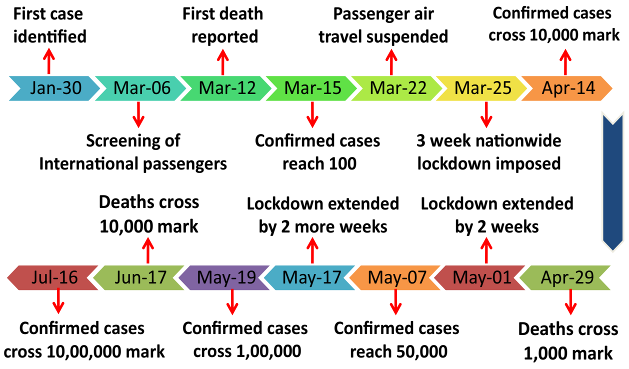
First case identified
First death reported
Passenger air travel suspended
Confirmed cases cross 10,000 mark
Jan-30
Mar-06
Mar-12
Mar-15
Mar-22
Mar-25
Apr-14
Screening of International passengers
Confirmed cases reach 100
3 week nationwide lockdown imposed
Deaths cross 10,000 mark
Lockdown extended by 2 more weeks
Lockdown extended by 2 weeks
Jul-16
Jun-17
May-19
May-17
May-07
May-01
Apr-29
Confirmed cases cross 10,00,000 mark
Confirmed cases cross 1,00,000
Confirmed cases reach 50,000
Deaths cross 1,000 mark

Fig. 3 A brief timeline of the COVID-19 pandemic in India, depicting the milestones with respect to the administrative measures taken, and the pandemic statistics

The Indian Government took several countermeasures initially to prevent the spread of the disease and to equip the hospitals and administration in order to be better prepared to deal with the pandemic. Beginning with the thermal screening of the passengers at airports, travel restrictions both domestically and internationally, closure of schools, workplaces, places of worship, etc., encouraging people to practice social distancing and wearing mask and a one day voluntary nationwide shutdown (https://www.mygov.in/covid-19/, Sarkar et al. 2020).

More stringent measures were enforced subsequently, starting with the imposition of a 3-week-long nationwide lockdown on March 25, 2020, which was progressively extended quadruple times each with a duration of 2 weeks (Sardar et al. 2020). A brief timeline of the COVID-19 pandemic in India is shown in Fig. 3. Being the second largest populous country of the world (inhabited by 1.39 billion people), and the fact that the pandemic is yet to slow down its spread, it becomes extremely important to know how the trajectory of the disease is likely to evolve in the country. There are apprehensions that the pandemic could potentially lead to a high surge in number of infected people and the deceased, and transform into a catastrophe, if the spread is not mitigated and effective control measures are not implemented. Thus, the prediction of the ensuing course of the disease is very important as the outbreak further unfolds in the near future. Mathematical modeling of the pandemic is thus essential as the forecasts based on such models would prove immensely useful to policy makers, administration, and health-care personnel and help them to formulate various strategies and be in a state of preparedness to deal with eventualities that may be inevitable. Modeling and forecasting of the pandemic coupled with the strict following of the guidelines are extremely important until a dedicated vaccine is developed for the disease.

Several researchers have modeled the COVID-19 pandemic and forecasted its future evolution. For instance, among the first works in this regard Lin et al. (2020) used an extension of the Susceptible-Exposed-Infectious-Removed (SEIR) compartmental model to account for the zoonotic origin of the COVID-19, individual response, governmental intervention and the emigration of the people in the early stages of the pandemic in Wuhan, China. A similar line of approach has been followed in Giordano et al. (2020) by Giordano et. al., their model SIDARTHE is also an extension of the SEIR model and includes other classes of people like diagnosed, ailing, threatened and healed. The SIDARTHE model has been used to model the COVID-19 spread in Italy. Anastassopoulou et al. (2020) estimated the transmission rate (R_0), case-fatality rate and recovery rate based on an SIDR model for Hubei province, China, from January 11, 2020–February 10, 2020 and also provided a 3-week-long forecast of these epidemiological parameters. The transmissibility of the COVID-19 from super spreaders (an individual or a mass gathering) has been modeled in Ndairou et al. (2020). Hellewell et al. (2020) have modeled the effectivenss of isolation of infected patients and contact tracing on the COVID-19 spread. Their simulation results suggest that an extensive contact tracing and isolation of suspected cases is very likely to suppress the pandemic within few months, if the infections start spreading after the onset of symptoms. Eikenberry et al. (2020) have modeled the effectiveness of using face masks on the spread of the virus. Their results demonstrate that a broad use of face masks can prevent the widespread transmission of the disease and thereby reduce hospitalization of patients and fatalities. Torrealba-Rodriguez et al. (2020) have modeled the spread of the COVID-19 outbreak in Mexico using the Gompertz and Logistic models, and the machine learning approach based on the artificial neural network. Furthermore, the authors used the model inversions to extrapolate the unfolding of the disease for a duration of 1 week. In Saba and Elsheikh (2020), the authors have used autoregressive integrated moving average (ARIMA) and nonlinear autoregressive artificial neural networks (NARANN)-based approaches to analyze the prevalence of the pandemic in the African country of Egypt, using the data obtained from the Egyptian ministry of health (MoH). The findings of the study reveal that NARANN is statistically better in modeling the behavior of the pandemic for the Egyptian data. The authors in Yousaf et al. (2020) have forecasted the number of cases, number of recoveries and the number of deaths for a one month period for Pakistan using ARIMA models and provided certain guidelines in order to contain the spread of the virus. The modeling and provision of future projections in Nigeria, Saudi Arabia, Brazil, Canada, USA, Japan and Australia have been undertaken in (Ayinde et al. 2020; Alzahrani et al. 2020; Ribeiro et al. 2020; Chimmula and Zhang 2020; Velásquez and Lara 2020; Kuniya 2020; Chang et al. 2020).

Modeling and forecasting of the COVID-19 pandemic in India has been attempted by Sarkar et al. (2020), using $SARII_qS_q$ model, which is a modified and a more general form of the SEIR model and takes into account the asymptomatic disease carriers (A), infected individuals who are isolated (I_q) and susceptible individuals who are under quarantine (S_q), besides taking the susceptible (S), recovered (R) and infected (I) persons into consideration. The authors have examined the reproduction rate (R_0) of the spread and how it varies with administrative measures like lockdowns. They have also simulated the lifecycle of the disease and obtained its possible date of termination. However, owing to the limited data used in the analysis, the findings of the work are not consistent with the prevailing scenario. For instance, their simulation results were suggestive that the pandemic would terminate on July 26, 2020; however, as on this date, it is still far from being over, and in fact, as mentioned earlier, India is currently witnessing a daily increase of about 50,000 cases. A short-term forecast (May 1, 2020–May 22, 2020) of COVID-19 cases in India has been done in Malavika et al. (2020), based on an SIR model and a logistic growth model, with the latter been found to be more effective than the former. The 3-week projections of their analysis, as compared to the actual data, however, turned out to be significantly lower. The study also reveals that the lockdowns that were enforced in the country to suppress the spread of the disease did not had any statistically significant effect on the mitigation of the transmission. Tomar and Gupta (2020) have used long short-term memory (LSTM) technique and the classical curve fitting for modeling and forecasting COVID-19 in India using a very limited data.

In this paper, we attempt to model the emergence of the COVID-19 pandemic in India and to obtain a one month long forecast based on the developed model(s). We model and forecast: (1) total number of diagnosed cases, (2) total number of recoveries and (3) total number of deaths, based on the publicly available data from March 1, 2020 to July 25, 2020. In addition to this, we also forecast the total number of active cases based on the forecasts obtained for the other three categories. Knowing that the pandemic in India has not shown any signs of slowing down yet and is still monotonously increasing, we assume that the trend will prevail in the upcoming month. Our forecasts therefore represent the worst case numbers of the likely infected cases, recoveries and casualties that can be anticipated. The rest of the paper is organized as follows: In the next section, the rationale for modeling and prediction of COVID-19 is first presented, followed by the description of the data sources and finally the mathematical preliminaries of the modeling technique are detailed. In Sect. 3, the results of the modeling and forecasting are provided and discussed briefly, finally the paper is concluded in Sect. 4.

Modeling and forecasting of COVID-19 in India

In this section, we mathematically model the progression of the COVID-19 pandemic in India from March 01, 2020 till July 25, 2020. We select a particular model within a certain class of models based on the statistical properties viz. root-mean-squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), coefficient of determination (R^2), and then attempt to predict the future course of the disease for a period of one month, i.e., upto August 24, 2020. The modeling and forecasting of the COVID-19 pandemic is formulated as a typical univariate time-series problem using the autoregressive integrated moving average (ARIMA) technique, wherein it is assumed that the current or future values of the diagnosed cases/recoveries/deaths are functions of their lagged (past) values. ARIMA models are typically used to model and forecast processes that yield a time-series as output, and have been used in varied areas ranging from weather forecasting (Wanishsakpong and Owusu 2020), transportation forecasting (Ediger and Akar 2007), fuel energy demand forecasting (Andreoni and Postorino 2006), milk production forecasting (Taye et al. 2020), groundwater level forecasting (Abuamra et al. 2020) to Stock price prediction (Ariyo et al. 2014).

In order to develop the model, the available datasets are divided into two subsets: (1) Training data (90%), and (2) Validation data (10%). The training data is used to determine the unknown model parameters, and the validation data is used to assess the forecasting capabilities of the developed models. A total of 29 models are obtained among which only one model is finally chosen that has better statistical characteristics than the others. The selected model is then used for forecasting month ahead projections from July 26, 2020, to August 24, 2020, for the total number of diagnosed cases, total number of recoveries and total number of deaths. Once these forecasts are obtained, the forecast for total number of active cases is obtained using:

$$N_{ac} = N_{dc} - N_r - N_d, \qquad (1)$$

where, N_{ac} is the total number of active cases, N_{dc} is the total number of diagnosed cases, N_r is the total number of recoveries, and, N_d is the total number of deaths.

Data source

The data utilized in this study has been obtained from several reliable sources that include the Government of India's Ministry of Health and Family Welfare (MoHFW) (https://www.mohfw.gov.in/), the website worldometer (https://www.worldometers.info/coronavirus/) and the website covidindia (https://covidindia.org/). The data has been collected from March 1, 2020, till July 25, 2020, and pertains to the cumulative number of diagnosed (infected) cases, cumulative number of recovered patients, and cumulative number of the deceased patients. The cumulative number of active cases was determined using Eq. 1. The data was preprocessed using the MATLAB environment, and any anomalies, if present, were accordingly verified and corrected. The data is shown in Fig. 4.

Mathematical preliminaries

In order to make this paper self-contained, a brief introduction to time-series analysis and forecasting using ARIMA(p,d,q) model is essential, and the same is presented here. For an in-depth coverage of these topics, the interested readers can refer to the classical books (Box et al. 2011; Montgomery et al. 2015).

A time-series is a collection of data-points which are measured or observed after successive fixed time durations, for e.g., hourly, daily, weekly, monthly or annually (Box et al. 2011). Time series is used to model and forecast various phenomena ranging from epidemics/pandemics, weather, stock markets, transportation, etc. (Yousaf et al. 2020; Andreoni and Postorino 2006; Ariyo et al. 2014). There are several models that can approximate their behavior and forecast their future evolution, for instance, exponential smoothing method, ARIMA(p,d,q), artificial neural network, logistic regression, etc., (Torrealba-Rodriguez

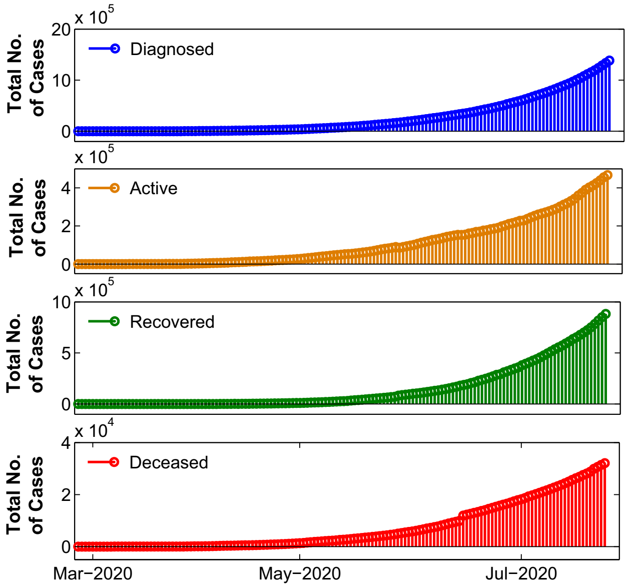

Fig. 4 COVID-19 cumulative statistics in India, from March 01, 2020 to July 25, 2020. The data has been compiled from (https://www.worldometers.info/coronavirus/, https://www.mohfw.gov.in/, https://covidindia.org/)

et al. 2020; Malavika et al. 2020; Montgomery et al. 2015). ARIMA(p,d,q), however, remains by far the most widely used.

The ARIMA(p,d,q) model comprises of the autoregressive part i.e., AR(p), and the moving average part i.e., MA(q), which have degrees p and q, respectively. The parameter d represents the order of differencing that is needed to stationarize the time series. The optimum degrees, p, d and q, of the ARIMA(p,d,q) model can be determined using Akaike information criterion (AIC), corrected AIC, Bayesian Information Criterion (BIC) or from the autocorrelation function(ACF) plots and the partial autocorrelation function (PACF) plots (Box et al. 2011; Montgomery et al. 2015; Yousaf et al. 2020). Besides these three hyper-parameters, the ARIMA(p,d,q) models have certain unknown coefficients that are usually determined using the maximum likelihood estimation or the least square estimation techniques (Box et al. 2011; Montgomery et al. 2015).

A univariate discrete time series is often represented as:

$$Y = \{Y_t : t \in \mathbb{Z}^+\}, \tag{2}$$

which can be written in expanded form as:

$$Y = \{Y_1, Y_2, Y_3, ...Y_T\}, \tag{3}$$

where, T is the total number of data-points in the time series.

In an ARMA(p,q) model, any general term of a non-seasonal and stationary time-series, $Y = \{Y_t : t \in \mathbb{Z}^+\}$, is assumed to be linearly dependent on the previous terms of the series, i.e.,

$$\begin{aligned} Y_t =& \epsilon_t + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} \\ &+ \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + ... + \beta_q \epsilon_{t-q}, \end{aligned} \tag{4}$$

where the $\alpha's$ represent the unknown coefficients in the AR(p) part of the model, and the $\beta's$ are the unknown coefficients of the MA(q) part of the ARMA(p,q) model, respectively. The $\epsilon's$ denote the error terms that are generally assumed to be normally distributed white noise signals with zero mean and a finite variance $\sigma_w^2 > 0$, i.e., $\epsilon_t \approx \epsilon_n(0, \sigma_w^2)$ (Box et al. 2011; Montgomery et al. 2015).

In compact form, the ARMA(p,q) model assumes the form:

$$Y_t = \epsilon_t + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \beta_j \epsilon_{t-j}, \tag{5}$$

The ARMA(p,q) model is often specified in an alternate form, that requires the use of the backshift operator or lag operator, that is defined as:

$$B^k Z_t = Z_{t-k}, \tag{6}$$

In terms of the backshift operator, Eq. 3 can be written as:

$$\begin{aligned} Y_t &= \epsilon_t + \alpha_1 BY_t + \alpha_2 B^2 Y_t + ... + \alpha_p B^p Y_t \\ &+ \beta_1 B\epsilon_t + \beta_2 B^2 \epsilon_t + ... + \beta_q B^q \epsilon_t. \end{aligned} \tag{7}$$

This can be put as:

$$(1 - \alpha_1 B + \alpha_2 B^2 + ... + \alpha_p B^p) Y_t = (1 + \beta_1 B + \beta_2 B^2 + ... + \beta_q B^q) \epsilon_t. \tag{8}$$

In compact form, this can be written as:

$$\alpha(B) Y_t = \beta(B) \epsilon_t. \tag{9}$$

The ARMA(p,q) models approximate the time-series behavior only if the series is stationary, i.e., when the statistical properties of the series are independent of the time interval in which they are observed. The following definitions (Box et al. 2011; Montgomery et al. 2015) of stationarity are often used in connection with time series:

1. A time-series $Y = \{Y_t : t \in \mathbb{Z}^+\}$ is said to be strictly stationary, if the statistical properties of $(Y_{t_1}, Y_{t_2}, Y_{t_3} ... Y_{t_n})$ remain invariant under a time shift operation. In other words, if the statistical properties of $(Y_{t_1}, Y_{t_2}, Y_{t_3} ... Y_{t_n})$ are exactly same as those of $(Y_{t_1+\tau}, Y_{t_2+\tau}, Y_{t_3+\tau} ... Y_{t_n+\tau}) \quad \forall \tau$

   This is however a very strong condition and difficult to verify analytically for a time series.
2. A time-series $Y = \{Y_t : t \in \mathbb{Z}^+\}$ is said to be weakly stationary, if both the mean of the series, Y, and the covariance of the series terms, Y_t and Y_{t-m}, exhibit invariance with respect to time. More specifically, a time series is weakly stationary, if:

   (a) The mean of the time-series $Y = \{Y_t : t \in \mathbb{Z}^+\}$ is a constant, i.e., $E(Y) = \mu$ (a constant), and
   (b) The covariance between the terms Y_t and Y_{t-m} with a window length of m is only a function of m, i.e., $Cov(Y_t, Y_{t-m}) = \eta(m)$.

In practice, the weak stationarity of a time series is often visually assessed by the plot of time-series data. If the plot appears to fluctuate about a certain mean value, then the series is stationary. Similarly, if the autocorrelation functions of the time-series exhibit a decaying trend then the series is stationary. A non-stationary time-series cannot be modeled by ARMA(p,q) model directly, unless the series is first stationarized. A non-stationary series is often made stationary by using differencing, in which a transformation of the following form is applied on the series:

$$Y'_t = Y_t - Y_{t-1}. \tag{10}$$

The order of differencing (d) needed to stationarize a time-series, varies as per the nature of the series. The ARMA(p,q) model coupled with a prior differencing is collectively known as ARIMA(p,d,q) model. The parameter estimation (α's and β's) and the forecasts based on the developed models are usually performed using dedicated programming routines of the commercially available software packages like in MATLAB, Mathematica, Python, R and Stata. We have, in fact, used MATLAB for the purpose here.

Application to indian COVID-19 data

A simple line plot of the total number of diagnosed cases is shown in Fig. 5, along with the first-order differenced, and the second-order differenced time series of the same data. The data is normalized to aid in the visualization of the trends in the series, as their ranges are different. The monotonously increasing trend of the actual data and the first-order differenced data implies the non-stationarity of the time series. The second-order differenced series, is however, stationary as its fluctuates about a mean value. Thus, to model this time series using ARMA(p,q), and to obtain forecasts based on this, a prior differencing of atleast a second order is necessary. We have thereby chosen two values of $d = 2$ and 3, to model this time-series here.

In order to determine the likely values of the hyperparameters p and q, we resort to the ACF and PACF plots of the time-series, corresponding to the total number of diagnosed cases. The ACF and PACF plots are shown in Fig. 6. Since the ACF and PACF plots, neither show a gradual damping, nor a cut-off at a single value of lag. We use several arbitrary values for the parameter, $p = 2, 5, 7, 8, 9, 14$ and 15, and, the parameter, $q = 2, 3, 5$ and 7. An iterative procedure is thereby used for various combinations of p, d and q to determine the model that exhibits better statistical properties and has superior forecasting capabilities. The statistical metrics of root-mean-square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and the coefficient of determination (R^2) are used for the model validation. The statistical parameters are determined using the following expressions:

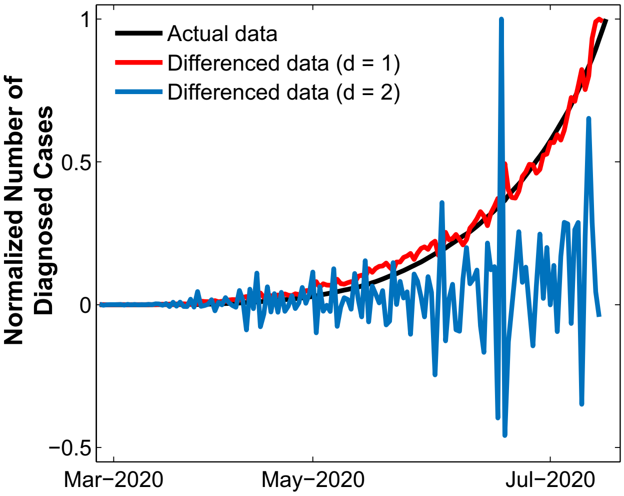

Fig. 5 Normalized plots of the cumulative number of diagnosed cases, first-order differenced data and the second-order differenced data. The original series and the first-order differenced series is non-stationary, while the second-order differenced series is stationary

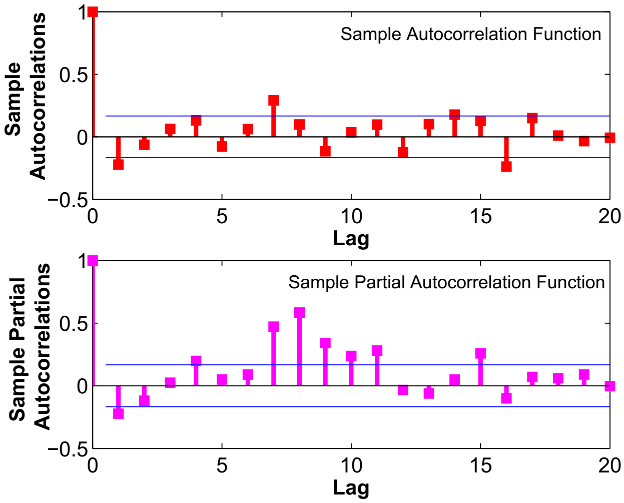

Fig. 6 Autocorrelation function (ACF) plot, and the partial autocorrelation function (PACF) plot of the second-order differenced series for the total number of diagnosed cases

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2}, \tag{11}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|Y_j - \hat{Y}_j|, \tag{12}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{j=1}^{n}\frac{|Y_j - \hat{Y}_j|}{|Y_j|}, \tag{13}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2}{\sum_{j=1}^{n}(Y_j - \bar{Y})^2}, \tag{14}$$

where Y_j is the actual value, $\hat{Y}_j$ is the predicted value, $\bar{Y}$ is the time-series mean and n is the number of observations taken.

The results of the iterations for the validation data set are depicted in Table 1. From this table, it can be clearly observed that the ARIMA(7,2,2) model, has the minimum values of the error metrics, and maximum value of the coefficient of determination, and therefore is the optimum

Table 1 Statistical metrics of the various ARIMA(p,d,q) models

S. no.	ARIMA(p,d,q) model	RMSE	MAE	MAPE	R^2
01	ARIMA(2,2,2)	597.7	395.1	0.428	0.99997
02	ARIMA(2,2,3)	551.4	384.98	0.36986	0.99997
03	ARIMA(2,2,5)	542.68	368.57	0.29912	0.99998
04	ARIMA(2,2,7)	552.18	402.06	0.65024	0.99997
05	ARIMA(5,2,2)	534.29	390.61	0.59157	0.99998
06	ARIMA(5,2,3)	538.31	388.46	0.49523	0.99998
07	ARIMA(7,2,2)	457.61	330.79	0.2471	0.99998
08	ARIMA(7,2,3)	481.5	341.27	0.34807	0.99998
09	ARIMA(7,2,5)	473.14	345.13	0.3912	0.99998
10	ARIMA(8,2,2)	510.71	358.26	0.65241	0.99998
11	ARIMA(8,2,3)	481.48	341.38	0.35068	0.99998
12	ARIMA(8,2,5)	480.48	342.45	0.45999	0.99998
13	ARIMA(9,2,2)	507.76	355.99	0.67046	0.99998
14	ARIMA(9,2,3)	481.33	342.28	0.34025	0.99998
15	ARIMA(9,2,5)	472.24	340.38	0.45992	0.99998
16	ARIMA(14,2,2)	480.42	341.09	0.67289	0.99998
17	ARIMA(15,2,2)	514.69	357.46	0.62912	0.99998
18	ARIMA(2,3,2)	558.4	394.53	0.44165	0.99997
19	ARIMA(2,3,5)	545.43	383.91	0.37512	0.99997
20	ARIMA(2,3,7)	514.46	397.19	0.47946	0.99998
21	ARIMA(5,3,2)	523.06	374.22	0.3845	0.99998
22	ARIMA(5,3,5)	544.49	354.65	0.26738	0.99997
23	ARIMA(5,3,7)	472.74	333.65	0.39325	0.99998
24	ARIMA(7,3,2)	511.03	358.35	0.46553	0.99998
25	ARIMA(7,3,5)	472.91	344.74	0.35958	0.99998
26	ARIMA(8,3,2)	507.77	359.32	0.4316	0.99998
27	ARIMA(8,3,5)	472.53	343.8	0.3787	0.99998
28	ARIMA(14,3,2)	463.39	340.99	0.54049	0.99998
29	ARIMA(15,3,2)	468.69	342.67	0.39887	0.99998

model. We assume that for the other datasets similar inferences are true. It is important to mention that the missing combinations of the p, d and q values turned out to unstable and hence it was not possible to establish their statistical parameters.

Results and discussions

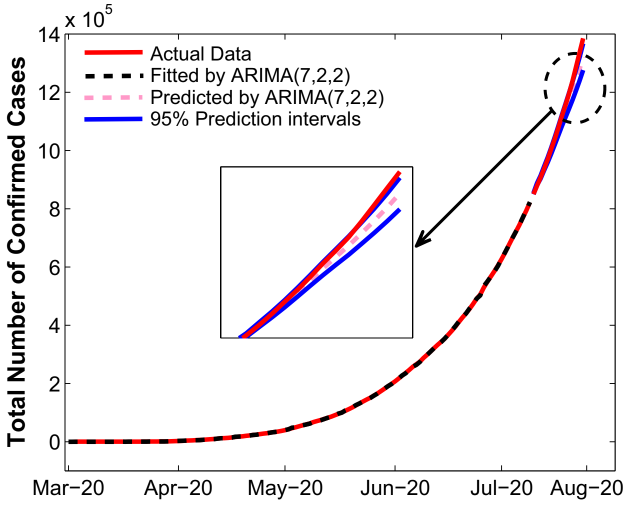

Fig. 7 A comparison of the actual data pertaining to the cumulative number of diagnosed cases, and the data fitted by and predicted by the ARIMA(p,d,q) model

The modeling and predicting capabilities of the ARIMA(7,2,2) models for the total number of diagnosed cases, total number of recoveries and the total number of deaths are illustrated in Figs. 7, 8 and 9. It has to be recalled that we had partitioned the available data from March 01, 2020, to July 25, 2020, into two groups: the first, for training purposes (90%), i.e., for model development, and the second, for validation purposes (10%), i.e., for assessing the predicting capabilities of the developed models. From Figs. 7, 8 and 9, it is clear that ARIMA(7,2,2) models have very good fitting (with respect to training data) and predicting (with respect to validation data) capabilities for all the three categories of datasets. The values predicted by the ARIMA(7,2,2) models are quite close to the actual values. The small deviations between the predicted values and the actual values have been quantified using the statistical metrics of RMSE, MAE, MAPE and R^2, and are provided in Table 1, for the total number of diagnosed cases.

The one-month ahead forecasts using the ARIMA(7,2,2) models, for the total number of diagnosed cases, total number of recoveries and the total number of deaths are shown

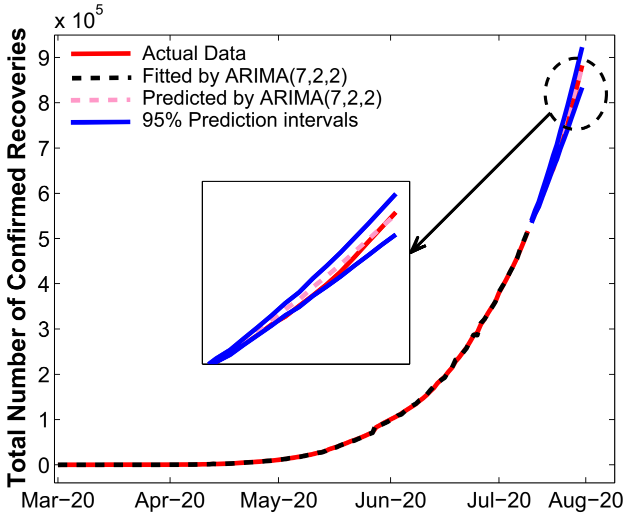

Fig. 8 A comparison of the actual data pertaining to the cumulative number of recoveries, and the data fitted by and predicted by the ARIMA(p,d,q) model

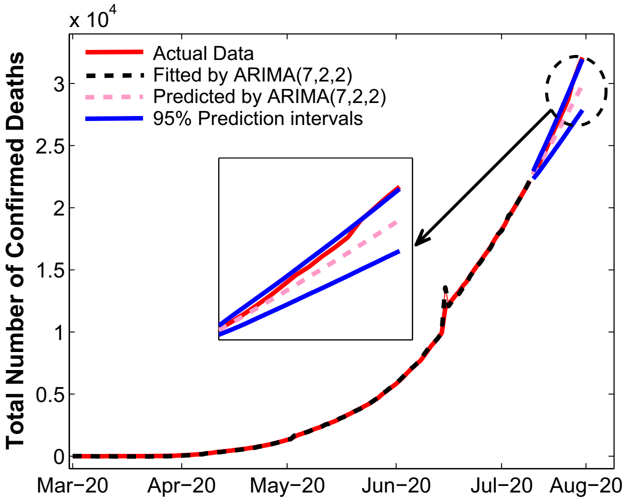

Fig. 9 A comparison of the actual data pertaining to the cumulative number of deaths, and the data fitted by and predicted by the ARIMA(p,d,q) model

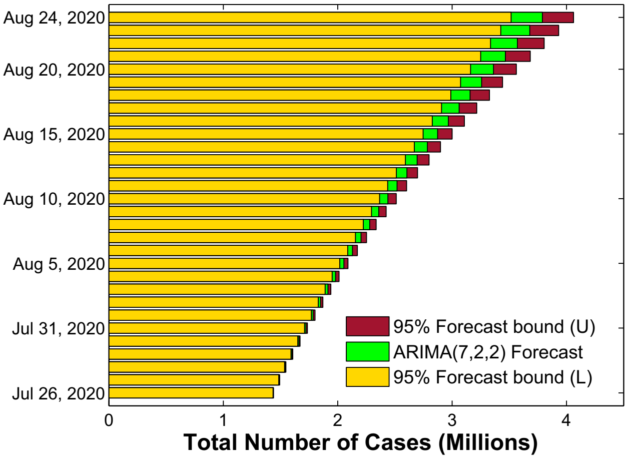

Fig. 10 A one-month forecast (July 26, 2020–August 24, 2020), along with the 95% confidence limits, of the cumulative number of diagnosed cases using ARIMA(7,2,2) model

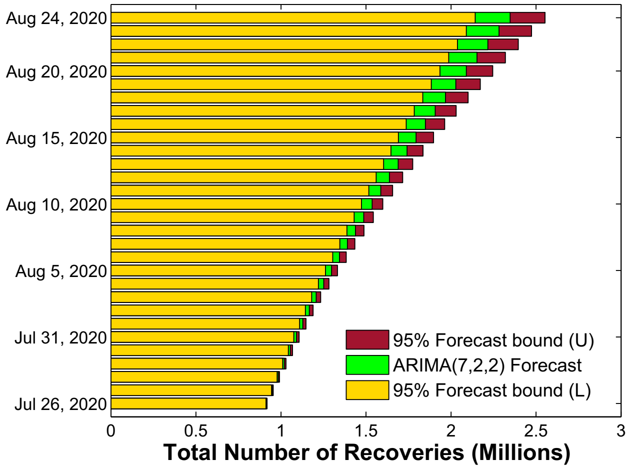

Fig. 11 A one-month forecast (July 26, 2020–August 24, 2020), along with the 95% confidence limits, of the cumulative number of recoveries using ARIMA(7,2,2) model

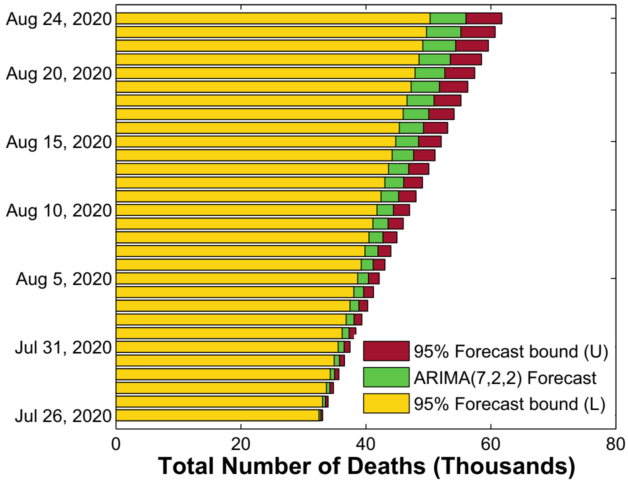

Fig. 12 A one-month forecast (July 26, 2020–August 24, 2020), along with the 95% confidence limits, of the cumulative number of deaths using ARIMA(7,2,2) model

in Figs. 10, 11 and 12, along with their 95% confidence intervals. The absolute values are provided in Table 2, along with the forecast for the total number of active cases. Our forecasts predict that by August 24, 2020, the expected number of cumulative diagnosed cases would increase nearly threefold from now, and surge to 3800,989, the number of recoveries would reach 2110,697 and the cumulative number of deaths would mount to 56,150. India is likely to cross the two million diagnosed cases mark on August 5, 2020, and the three million cases on August 17, 2020. The cumulative recoveries are expected to breach the two million mark on August 22, 2020, and the cumulative deaths could hit the fifty thousand mark on August 17, 2020. The daily increment in cumulative diagnosed would be nearly 110,182, for recoveries it will be 45,914, and for deaths the value would be 874. These forecasts would also correspond to a recovery rate of 55.53% and case-fatality rate of 1.47%. In comparison with the current values of 64.53% and 2.23%, both the recovery rate and the case-fatality rate would be lower, while a lower case fatality rate is desirable; however, a lower recovery rate would be a matter of concern. With regard to the rate of COVID-19 spread in the upcoming month, it is found that the cumulative diagnosed cases would escalate by 63.54%, which is lower than the current value of 65% and therefore would be promising.

Table 2 One-month ahead forecast (July 26, 2020–August 24, 2020) of total number of diagnosed cases, total number of recoveries, total number of deaths and total number of active cases in India

S. no.	Total no. of diagnosed cases	Total no. of recoveries	Total no. of deaths	Total no. of active cases
1	1436,970 (1439,742, 1434,197)	916,949 (920,450, 913,448)	32,809 (33,107, 32,510)	487,212
2	1488,550 (1493,153, 1483,946)	952,696 (958,451, 946,942)	33,569 (34,019, 33,118)	502,285
3	1542,049 (1548,455, 1535,643)	989,676 (998,366, 980,987)	34,316 (34,895, 33,737)	518,057
4	1599,989 (1608,489, 1591,490)	1027,654 (1039,853, 1015,454)	35,076 (35,776, 34,375)	537,259
5	1661,463 (1672,585, 1650,341)	1065,996 (1082,302, 1049,691)	35,823 (36,642, 35,004)	559,644
6	1723,752 (1737,844, 1709,660)	1103,893 (1125,261, 1082,524)	36,573 (37,524, 35,623)	583,286
7	1786,352 (1804,051, 1768,652)	1141,193 (1167,754, 1114,633)	37,331 (38,417, 36,244)	607,828
8	1850,560 (1873,047, 1828,074)	1179,167 (1211,860, 1146,475)	38,093 (39,321, 36,865)	633,300
9	1916,205 (1944,187, 1888,223)	1218,547 (1257,883, 1179,212)	38,862 (40,234, 37,491)	658,796
10	1984,113 (2017,926, 1950,300)	1259,177 (1305,608, 1212,746)	39,635 (41,152, 38,118)	685,301
11	2055,690 (2095,741, 2015,638)	1300,182 (1354,122, 1246,242)	40,413 (42,078, 38,747)	715,095
12	2130,546 (2177,436, 2083,657)	1340,729 (1402,452, 1279,006)	41,195 (43,013, 39,377)	748,622
13	2206,770 (2261,133, 2152,406)	1380,691 (1450,816, 1310,567)	41,982 (43,955, 40,009)	784,097
14	2283,610 (2346,262, 2220,959)	1420,456 (1499,368, 1341,544)	42,775 (44,907, 40,643)	820,379
15	2361,733 (2433,697, 2289,770)	1460,953 (1549,155, 1372,752)	43,572 (45,866, 41,279)	857,208
16	2441,719 (2523,874, 2359,564)	1502,730 (1600,620, 1404,841)	44,375 (46,834, 41,917)	894,614
17	2524,282 (2617,255, 2431,308)	1545,514 (1653,395, 1437,632)	45,183 (47,810, 42,556)	933,585
18	2610,185 (2714,570, 2505,799)	1588,414 (1706,629, 1470,200)	45,996 (48,793, 43,198)	975,775
19	2699,103 (2815,615, 2582,592)	1630,626 (1759,429, 1501,824)	46,814 (49,786, 43,842)	1021,663
20	2789,688 (2919,145, 2660,231)	1672,212 (1812,007, 1532,417)	47,637 (50,786, 44,488)	1069,839
21	2881,174 (3024,520, 2737,827)	1713,897 (1865,060, 1562,735)	48,465 (51,794, 45,136)	1118,812
22	2973,945 (3132,225, 2815,665)	1756,556 (1919,453, 1593,660)	49,299 (52,811, 45,786)	1168,090
23	3068,765 (3242,950, 2894,580)	1800,472 (1975,416, 1625,527)	50,137 (53,836, 46,439)	1218,156
24	3166,347 (3357,248, 2975,445)	1845,102 (2032,324, 1657,881)	50,981 (54,868, 47,094)	1270,264
25	3267,126 (3475,504, 3058,749)	1889,550 (2089,311, 1689,790)	51,830 (55,909, 47,750)	1325,746
26	3370,749 (3597,440, 3144,058)	1933,237 (2145,792, 1720,683)	52,684 (56,957, 48,410)	1384,828
27	3476,193 (3722,148, 3230,238)	1976,444 (2202,116, 1750,773)	53,543 (58,014, 49,071)	1446,206
28	3582,794 (3849,065, 3316,523)	2020,040 (2259,157, 1780,923)	54,407 (59,078, 497,35)	1508,347
29	3690,807 (3978,498, 3403,115)	2064,783 (2317,638, 1811,928)	55,276 (60,151, 50,401)	1570,748
30	3800,989 (4111,155, 3490,823)	2110,697 (2377,543, 1843,852)	56,150 (61,231, 51,070)	1634,142

The values within the parantheses are the upper and lower 95% confidence limits

The forecasts of this study can be helpful to the authorities to put effective control efforts in place and to take timely actions to contain the spread of the pandemic. Furthermore, the authorities can equip themselves with sufficient number of hospital beds, ventilators, etc., and accordingly be well-prepared to deal with the overwhelming of the hospitals. Although the forecasts reported here are based on the actual pandemic data, however, it is to be noted that these forecasts are subjected to many influencing factors and thereby the actual numbers could be different than the ones reported here. The most important factors would be the availability of vaccine, testing rates, adherence to measures like social distancing, hand-washing, sanitizing and wearing of face masks. Besides these, certain other factors that include, the phased relaxations offered by the local bodies and Government, economic conditions of individuals and the country as a whole and mass migrations of the people would also affect the numbers reported here.

Conclusion

In this paper, the authors have modeled the current spread of the COVID-19 pandemic in India using an autoregressive integrated moving average (ARIMA) model and predicted its likely evolution for the upcoming one month period. The forecasts for the total number of diagnosed cases, active cases, recoveries and deaths are made based on an optimal ARIMA(7,2,2) model that has been selected within a class of 29 ARIMA(p,d,q) models based on various statistical parameters that provide a measure of the fit between the model outcomes and the actual results. The forecasting results reveal that the pandemic is likely to spread at a much faster rate, while the recoveries are going to slow down and the fatality ratio is likely to reduce. The forecasted results are worrying and suggest that unless new control measures are devised and implemented, and the established guidelines are strictly followed, the pandemic has the potential to turn devastating.

Author contributions SG: involved in conceptualization, investigation, methodology, formal analysis, supervision, software, writing–original draft, writing–review and editing. ON: involved in conceptualization, validation, writing–review and editing.

Funding The authors received no financial support for the research, authorship and publication of this article.

Data availability Available on request.

Code availability Available on request.

Compliance with ethical standards

Conflict of interest The authors declare that they have no known conflict of interest.

References

Abuamra IA, Maghari AY, Abushawish HF (2020) Medium-term forecasts for salinity rates and groundwater levels. Model Earth Syst Environ 1–10

Alzahrani SI, Aljamaan IA, Al-Fakih EA (2020) Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using Arima prediction model under current public health interventions. J Infect Public Health

Anastassopoulou C, Russo L, Tsakris A, Siettos C (2020) Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS One 15(3):e0230405

Andreoni A, Postorino MN (2006) A multivariate arima model to forecast air transport demand. In: Proceedings of the Association for European Transport and Contributors, pp 1–14

Ariyo AA, Adewumi AO, Ayo CK (2014) Stock price prediction using the arima model. In: 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, pp 106–112. IEEE

Ayinde K, Lukman AF, Rauf RI, Alabi OO, Okon CE, Ayinde OE (2020) Modeling nigerian COVID-19 cases: a comparative analysis of models and estimators. Chaos Solitons Fractals 138:109911

Box GE, Jenkins GM, Reinsel GC (2011) Time series analysis: forecasting and control, vol 734. Wiley, Hoboken

Chang SL, Harding N, Zachreson C, Cliff OM, Prokopenko M (2020) Modelling transmission and control of the COVID-19 pandemic in Australia. arXiv preprint arXiv:2003.10218

Chimmula VKR, Zhang L (2020) Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals. 109864

de Haas M, Faber R, Hamersma M (2020) How covid-19 and the dutch “intelligent lockdown” change activities, work and travel behaviour: Evidence from longitudinal data in the Netherlands. Transp Res Interdisciplinary Perspect. 100150

Di Renzo L, Gualtieri P, Pivari F, Soldati L, Attinà A, Cinelli G, Leggeri C, Caparello G, Barrea L, Scerbo F et al (2020) Eating habits and lifestyle changes during COVID-19 lockdown: an Italian survey. J Transl Med 18(1):1–15

Ediger VŞ, Akar S (2007) Arima forecasting of primary energy demand by fuel in Turkey. Energy Policy 35(3):1701–1708

Eikenberry SE, Mancuso M, Iboi E, Phan T, Eikenberry K, Kuang Y, Kostelich E, Gumel AB (2020) To mask or not to mask: modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. Infect Dis Model

Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, Colaneri M (2020) Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nat Med 1–6

Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, Munday JD, Kucharski AJ, Edmunds WJ, Sun F, et al (2020) Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Global Health

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet 395(10223):497–506

Kuniya T (2020) Prediction of the epidemic peak of coronavirus disease in Japan, 2020. J Clin Med 9(3):789

Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KS, Lau EH, Wong JY et al (2020) Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Eng J Med

Lin Q, Zhao S, Gao D, Lou Y, Yang S, Musa SS, Wang MH, Cai Y, Wang W, Yang L, et al (2020) A conceptual model for the outbreak of coronavirus disease 2019 (COVID-19) in Wuhan, China with individual reaction and governmental action. Int J Infect Dis

Malavika B, Marimuthu S, Joy M, Nadaraj A, Asirvatham ES, Jeyaseelan L (2020) Forecasting COVID-19 epidemic in India and high incidence states using sir and logistic growth models. Clin Epidemiol Global Health

Montgomery DC, Jennings CL, Kulahci M (2015) Introduction to time series analysis and forecasting. Wiley, Hoboken

Morawska L, Cao J (2020) Airborne transmission of sars-cov-2: the world should face the reality. Environ Int. 105730

Ndairou F, Area I, Nieto JJ, Torres DF (2020) Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. Chaos Solitons Fractals. 109846

Omori R, Mizumoto K, Chowell G (2020) Changes in testing rates could mask the novel coronavirus disease (COVID-19) growth rate. Int J Infect Dis

Ribeiro MHDM, da Silva RG, Mariani VC, dos Santos Coelho L (2020) Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. Chaos Solitons Fractals. 109853

Saba AI, Elsheikh AH (2020) Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. Process Saf Environ Protect

Saidan MN, Shbool MA, Arabeyyat OS, Al-Shihabi ST, Al Abdallat Y, Barghash MA, Saidan H (2020) Estimation of the probable outbreak size of novel coronavirus (COVID-19) in social gathering events and industrial activities. Int J Infect Dis

Sardar T, Nadim SS, Rana S, Chattopadhyay J (2020) Assessment of lockdown effect in some states and overall India: a predictive mathematical study on COVID-19 outbreak. Chaos Solitons Fractals. 110078

Sarkar K, Khajanchi S, Nieto JJ (2020) Modeling and forecasting the COVID-19 pandemic in India. Chaos Solitons Fractals. 110049

Sarmadi M, et al (2020) Association of COVID-19 global distribution and environmental and demographic factors: an updated three-month study. Environ Res. 109748

Su S, Wong G, Shi W, Liu J, Lai AC, Zhou J, Liu W, Bi Y, Gao GF (2016) Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol 24(6):490–502

Taye BA, Alene AA, Nega AK, Yirsaw BG (2020) Time series analysis of cow milk production at Andassa dairy farm, west Gojam zone, Amhara region, Ethiopia. Model Earth Syst Environ 1–9

Tomar A, Gupta N (2020) Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Sci Total Environ. 138762

Torrealba-Rodriguez O, Conde-Gutiérrez R, Hernández-Javier A (2020) Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. Chaos Solitons Fractals. 109946

Velásquez RMA, Lara JVM (2020) Forecast and evaluation of COVID-19 spreading in USA with reduced-space gaussian process regression. Chaos Solitons Fractals. 109924

Wanishsakpong W, Owusu BE (2020) Optimal time series model for forecasting monthly temperature in the southwestern region of Thailand. Model Earth Syst Environ 6(1):525–532

Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY et al (2020) A new coronavirus associated with human respiratory disease in China. Nature 579(7798):265–269

Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, Meng J, Zhu Z, Zhang Z, Wang J et al (2020) Genome composition and divergence of the novel coronavirus (2019-ncov) originating in China. Cell Host Microbe

Yousaf M, Zahir S, Riaz M, Hussain SM, Shah K (2020) Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan. Chaos Solitons Fractals. 109926

Zhang T, Wu Q, Zhang Z (2020) Probable pangolin origin of sars-cov-2 associated with the COVID-19 outbreak. Curr Biol

Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579(7798):270–273

Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R et al (2020) A novel coronavirus from patients with pneumonia in China, 2019. N Eng J Med

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.