Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Applied Soft Computing 122 (2022) 108806

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A comparative study for predictive monitoring of COVID-19 pandemic

Binish Fatimah [a,1], Priya Aggarwal [b,1], Pushpendra Singh [c,*], Anubha Gupta [d]

[a] Department of ECE, CMR Institute of Technology, Bengaluru, India
[b] Vehant Technologies Pvt. Ltd., Noida, India
[c] Department of ECE, National Institute of Technology Hamirpur, HP, India
[d] SBILab, Department of ECE, IIIT-Delhi, Delhi, India

ARTICLE INFO

Article history:
Received 10 July 2021
Received in revised form 2 January 2022
Accepted 31 March 2022
Available online 7 April 2022

Keywords:
COVID-19 modeling
Gaussian mixture model
Composite logistic growth function
SIRD model
ARIMA model
Dictionary learning model

ABSTRACT

COVID-19 pandemic caused by novel coronavirus (SARS-CoV-2) crippled the world economy and engendered irreparable damages to the lives and health of millions. To control the spread of the disease, it is important to make appropriate policy decisions at the right time. This can be facilitated by a robust mathematical model that can forecast the prevalence and incidence of COVID-19 with greater accuracy. This study presents an optimized ARIMA model to forecast COVID-19 cases. The proposed method first obtains a trend of the COVID-19 data using a low-pass Gaussian filter and then predicts/forecasts data using the ARIMA model. We benchmarked the optimized ARIMA model for 7-days and 14-days forecasting against five forecasting strategies used recently on the COVID-19 data. These include the auto-regressive integrated moving average (ARIMA) model, susceptible–infected–removed (SIR) model, composite Gaussian growth model, composite Logistic growth model, and dictionary learning-based model. We have considered the daily infected cases, cumulative death cases, and cumulative recovered cases of the COVID-19 data of the ten most affected countries in the world, including India, USA, UK, Russia, Brazil, Germany, France, Italy, Turkey, and Colombia. The proposed algorithm outperforms the existing models on the data of most of the countries considered in this study.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

In early December 2019, cases of the coronavirus disease (COVID-19) originated in Wuhan city by a Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1,2]. Within a short span of time, this virus quickly spread to a large population all over the world. It was declared an epidemic by the World Health Organization (WHO) on 11th March 2020. This disease is highly contagious and has infected millions of people globally. The number of deaths reported globally as of 15th May 2021 is more than 3.5 million. It has hugely affected economic activities [3] and plunged millions into poverty. In countries such as USA, Brazil, Italy, and India, the rapid increase in the number of cases caused tremendous stress on the health care system. Its spread has been contained to some extent in various countries by using partial and complete lock-downs, maintaining social distancing, and imposing quarantine for the infected people. For the timely implementation of these measures, a mathematical understanding of the future trend of the spread of disease is required. This can help the authorities announce control measures at an appropriate time. Thus, an accurate forecast of COVID-19 cases is extremely important to control its rapid spread and hence, ensure the safety of the general public.

Researchers across the world have proposed various data-driven methods to forecast COVID-19 data, which has been a difficult and challenging task [4,5]. Predicting or forecasting refers to estimating future cases on the basis of present and past data. It is carried out majorly using two popular approaches. The first approach includes compartmental models such as SIR, SIRD, SEIR models [6] and the second is based on time-series learning methods such as curve-fitting [7,8], autoregression [9,10], and deep learning on time-series data [11,12].

Compartmental models are the traditional methods of forecasting infectious diseases [13]. In these models, the spread of infectious diseases is simulated by stochastic differential equations that describe interactions between different compartments of the population (e.g. susceptible, infectious, and recovered). This approach majorly includes Susceptible–Infected–Removed (SIR) [14], Susceptible–Infected–Removed–Death (SIRD) [15,16] and Susceptible–Exposed–Infected–Removed (SEIR) models [17]. Hybrid models designed using compartmental models and deep learning frameworks have also been proposed recently [18]. Compartmental models are based on the assumption that the chance of an infected person to infect another susceptible person is

* Corresponding author.
E-mail addresses: binish.fatimah@gmail.com (B. Fatimah), priyaa@iiitd.ac.in (P. Aggarwal), pushpendrasingh@iitkalumni.org (P. Singh), anubha@iiitd.ac.in (A. Gupta).

[1] Equal contribution.

https://doi.org/10.1016/j.asoc.2022.108806
1568-4946/© 2022 Elsevier B.V. All rights reserved.

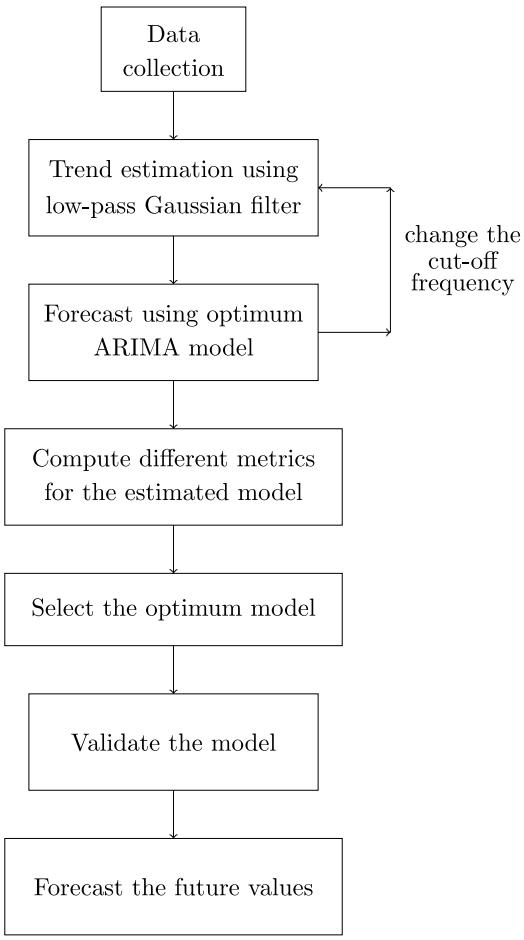


Fig. 1. Block diagram of the proposed methodology.

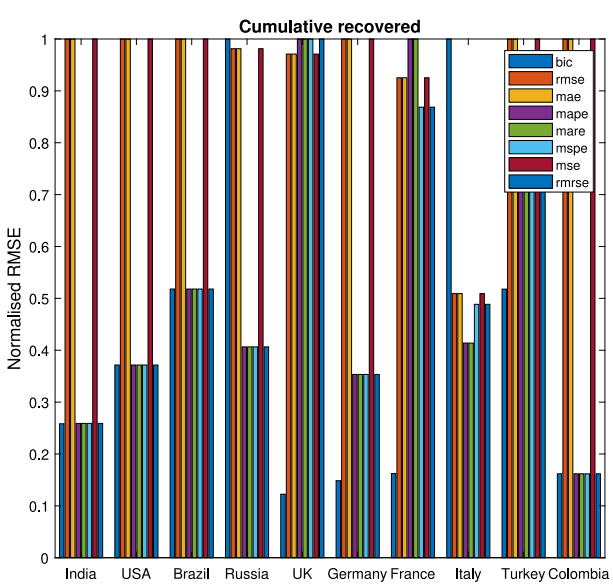


Fig. 2. Normalized RMSE values for cumulative recovered COVID-19 data, obtained when different metrics are used to select the optimum ARIMA model.

constant during the epidemic duration and also, every infected person has a constant chance to recover at any given time, which might not be true [19].

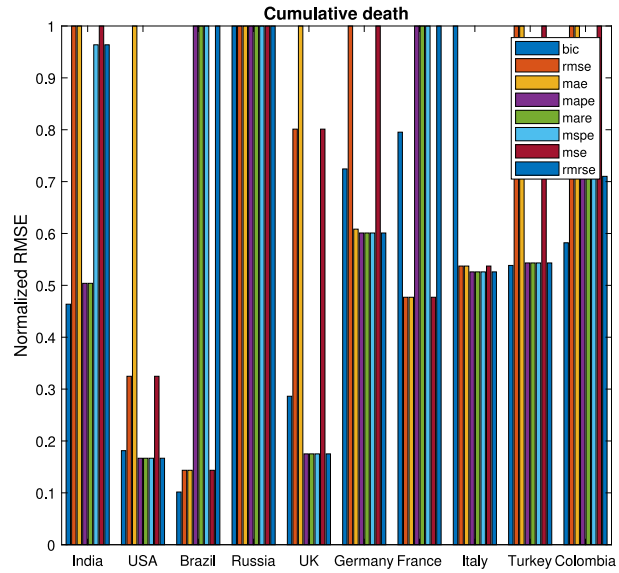


Fig. 3. Normalized RMSE values for the cumulative COVID-19 death data, obtained when different metrics are used to select the optimum ARIMA model.

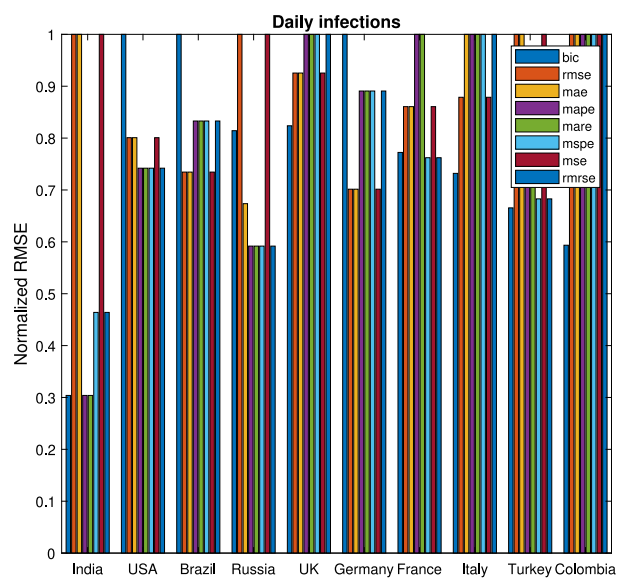


Fig. 4. Normalized RMSE values for the daily new COVID-19 infections data, obtained when different metrics are used to select the optimum ARIMA model.

Another popular approach is to fit the curves of certain shapes such as logistic [20] and Gaussian [21] to the available data and find parameters that yield optimal results with curve-fitting. Simple curve-fitting approaches typically support parameter estimation of a single wave characterized by a single peak throughout the epidemic duration. However, fitting the data with only one wave may be incorrect since, in general, there are several recurring waves that emerge and die throughout the epidemic duration [22]. To overcome this drawback, some recent works have decomposed the available data into multiple overlapping waves, where every single wave is a generalized growth model such as the logistic or the Gaussian growth models [7,8,14]. This is to note that both the compartmental and the curve fitting approaches are model-driven and require the estimation of parameters of some predefined mathematical model. Recently, a completely new approach has been applied to predict the spread of COVID-19. This includes sparse representation based on dictionary learning and

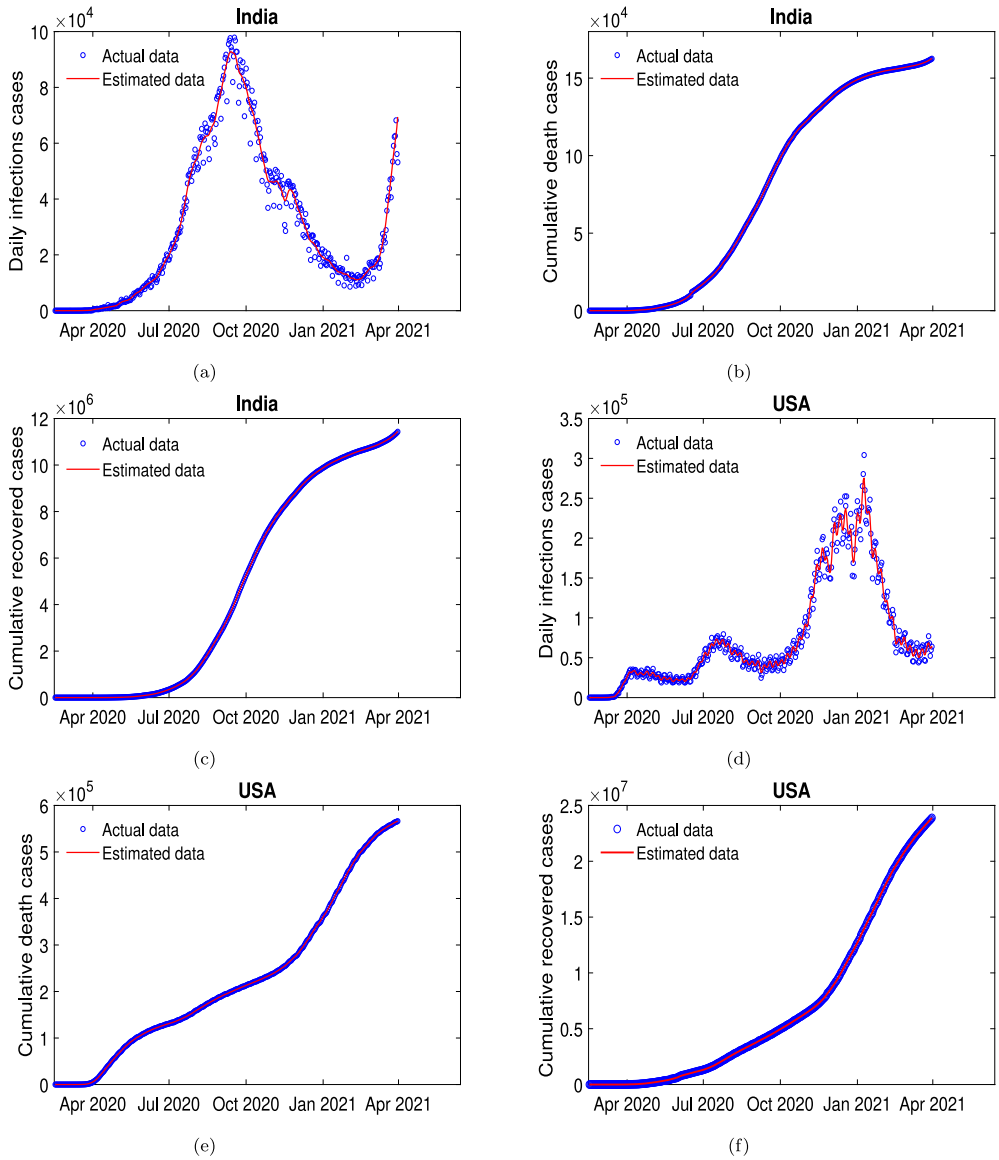


Fig. 5. Actual COVID-19 data and the estimated values as obtained using the proposed modeling scheme for 5(a) daily infections, 5(b) death cases, and 5(c) recovered cases for India, 5(d) daily infections, 5(e) death cases, and 5(f) recovered cases for USA.

Online Non-negative Matrix Factorization (ONMF) [23]. Modeling of signals through sparse representation has been very useful in different applications [24–26]. This modeling approach is totally model-free and does not assume any predefined mathematical model for prediction. In this method, evolution of COVID-19 reported cases is expressed as a sparse linear combination of the dictionary atoms. By recursively and progressively improving dictionary atoms to the most recent COVID-19 data, forecasting of cases is done using partial fitting [23]. Another widely used time-series approach is the Auto-Regressive Integrated Moving Average (ARIMA) modeling [27] because the data of many countries has an inherent non-stationary trend. Thus, the ARIMA model has been used widely by various authors for modeling and forecasting COVID-19 data [28–30]. To improve the performance of traditional ARIMA, Sharma et al. [31] used eigenvalue decomposition of the Hankel matrix to decompose the time-series into various stationary and non-stationary components. The decomposed signals were then modeled using ARIMA.

So far, several types of methods have been proposed for describing the time evolution of COVID-19 epidemic. Irrespective of the huge progress in proposing various methods for COVID-19 prediction, this research area is still nascent and requires a comparison of various prediction methods in detail. The literature search did not reveal any review of available models and thus, this work reviews various approaches mentioned in brief above. In addition, we compare the performance of these models by assessing the Root Mean squared error (RMSE) obtained for predicting the 7 days and 14 days future cases for the data of highly infected countries including India, USA, UK, Russia, Brazil, Germany, France, Italy, Turkey, and Colombia. We considered five different models including ARIMA, SIRD, composite Gaussian growth model, composite Logistic growth model, and ONMF model. The performance of the ARIMA model in predicting the short-term future is better than the other models for most of the cases as observed using the simulation studies. To further improve the results of the existing ARIMA model, we propose a modeling scheme by first filtering the data using a low-pass Gaussian filter to estimate the trend in the data. This low-frequency version of the data is then modeled using the optimized ARIMA models. Experimental results show that the

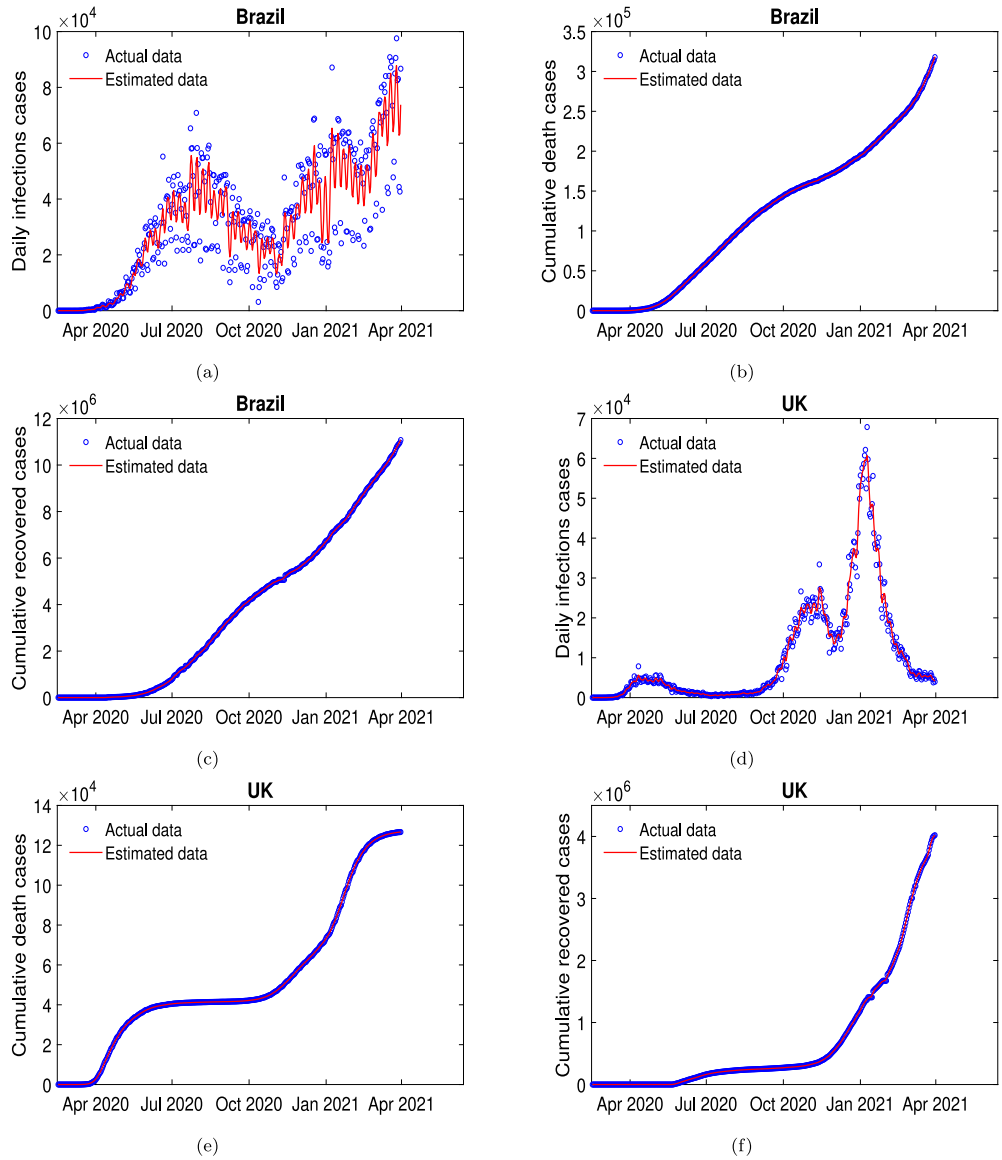


Fig. 6. Actual COVID-19 data and the estimated values as obtained using the proposed modeling scheme 6(a) daily infections, 6(b) death cases, 6(c) recovered cases for Brazil, 6(d) daily infections, 6(e) death cases, and 6(f) recovered cases for UK.

performance of the proposed optimized ARIMA model is superior to traditional ARIMA in most of cases because it is able to efficiently estimate the trend of the data.

The main contributions of this study are summarized as below:

- We have carried out a comprehensive review, comparison, and benchmarking of five popular data modeling methods, namely SIR, ARIMA, composite logistic growth model, Composite Gaussian model, and dictionary learning for forecasting and monitoring of COVID-19 pandemic.
- We have also proposed an optimized ARIMA model for predicting COVID-19 cases. This method provides minimum prediction error compared to the existing methods.
- We have reported 7-days and 14-days forecasting of the number of infected, recovered, and deaths of the COVID-19 data for the ten most affected counties. These models provide good prediction accuracy for the upcoming three weeks. However, the prediction accuracy declines gradually with the increase in prediction time.

This paper is organized as follows. In Section 2, we provide a brief overview of various forecasting methods available in the literature on COVID-19 data modeling. In Section 3, we provide a description of our proposed method. In Section 4, we present results. Finally, we discuss various results and conclude in Sections 5 and 6, respectively.

2. Existing methodologies

In this section, we discuss some of the existing modeling methods popularly used in the literature for modeling and forecasting COVID-19 data.

2.1. Susceptible-Infected-Removed (SIR) Models

SIR model [39] or its different variants such as the Susceptible–Infected–Recovered–Death (SIRD) model [32] or the Susceptible–Exposed–Infectious–Removed (SEIR) model [17] have been used to model the spread of diseases like dengue fever and malaria [40, 41]. Recently, various authors, [15,16,33,42], have used these methods for the modeling of data of COVID-19 prevalence. The

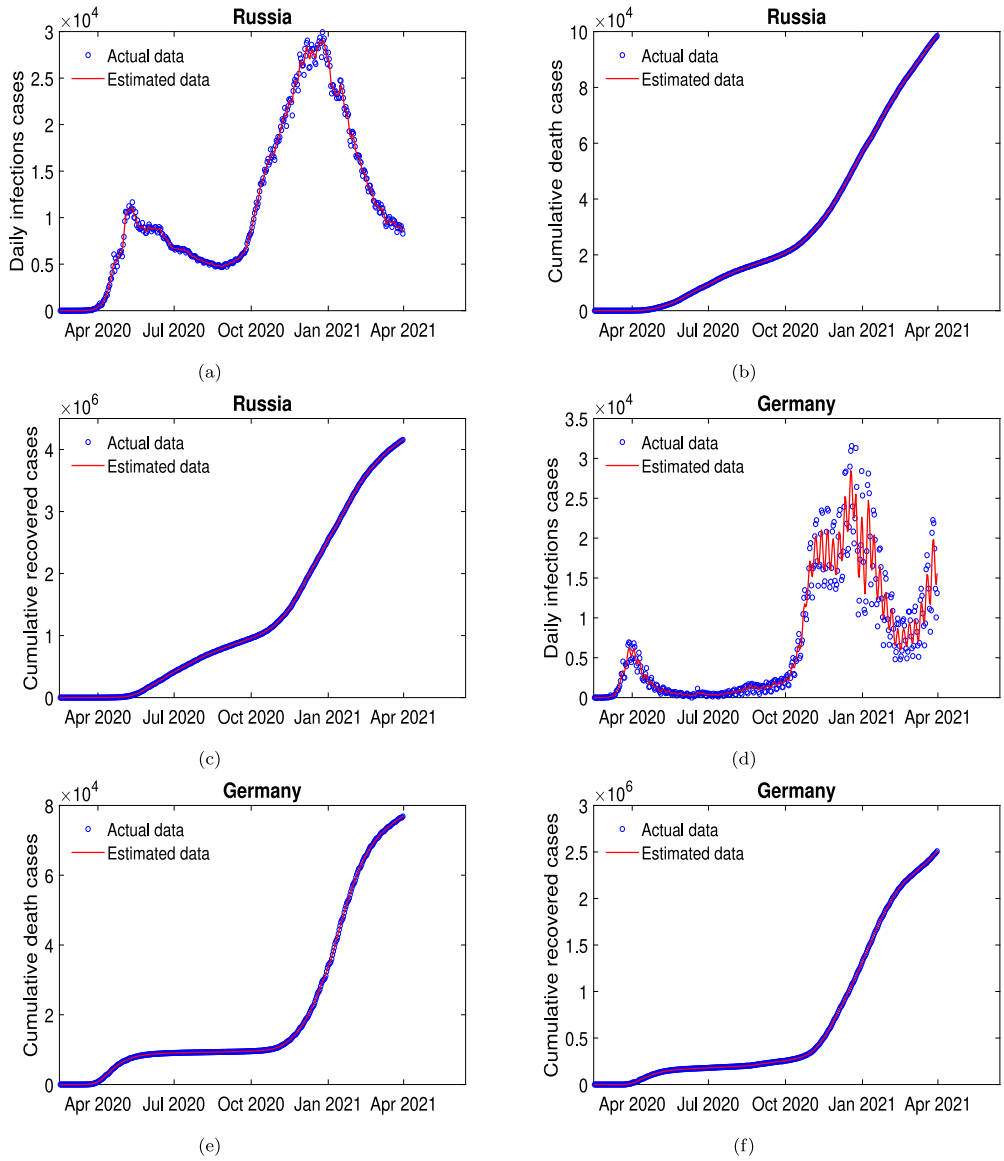


Fig. 7. Actual COVID-19 data and the estimated values as obtained using the proposed modeling scheme for 7(a) daily infections, 7(b) death cases, and 7(c) recovered cases for Russia, for 7(d) daily infections, 7(d) daily infections, 7(e) death cases, and 7(f) recovered cases for Germany.

traditional Susceptible–Infected–Recovered–Dead (SIRD) model [32] can be described using the following equations:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{-\beta IS}{N}, \tag{1a}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\beta SI}{N} - \gamma I - \mu I, \tag{1b}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I, \tag{1c}$$

$$\frac{\mathrm{d}D}{\mathrm{d}t} = \mu I, \tag{1d}$$

where $S(t)$, $I(t)$, $R(t)$, and $D(t)$ are the numbers of susceptible, infected, recovered, and death cases, respectively, β is the contact/infection rate (i.e., the average number of contacts per person per unit time) and γ is the recovery rate, i.e., $1/\gamma$ represents the average infectious period. Here, μ is the death rate. The initial values considered are $S(0) = S_0 \geq 0, I(0) = I_0 \geq 0, R(0) = R_0 \geq 0$, and $D(0) = D_0 \geq 0$. It is assumed that $S(t) + I(t) + R(t) + D(t) = N$, where N is a constant and refers to the total population size. An important feature of the SIRD model is the estimated reproduction number, $\mathfrak{R}_0 = \frac{\beta S}{N(\gamma+\mu)} > 1$. This number provides an indication about the spread of the disease as the number of susceptible cases getting infected from one infected person. If $\mathfrak{R}_0 > 1$, the number of cases are increasing, as in the start of an epidemic, $\mathfrak{R}_0 = 1$ indicates the disease is endemic, and $\mathfrak{R}_0 < 1$ indicates a decline in the number of cases.

In [42], SIR approach has been used to model the prevalence of COVID-19 data in China. In [15], a modified SIRD model is proposed to estimate COVID-19 data for five countries including India, USA, China, Italy, and France. This model considers the active, dead, and recovered cases simultaneously. It also considers the effect of quarantine and asymptomatic cases on the SIRD model that was otherwise not present in the traditional SIRD model. In [16], SIRD model is used on the COVID-19 data of Italy. The parameters of the proposed model were considered to be time-varying and were expressed as linear combinations of the basis functions. Sparse identification methodologies were used to obtain these functions from the given COVID-19 data. The non-convex identification problem of estimating the model parameters was handled by a one-dimensional grid search in the outer loop and using Lasso optimization in the inner step. In [34],

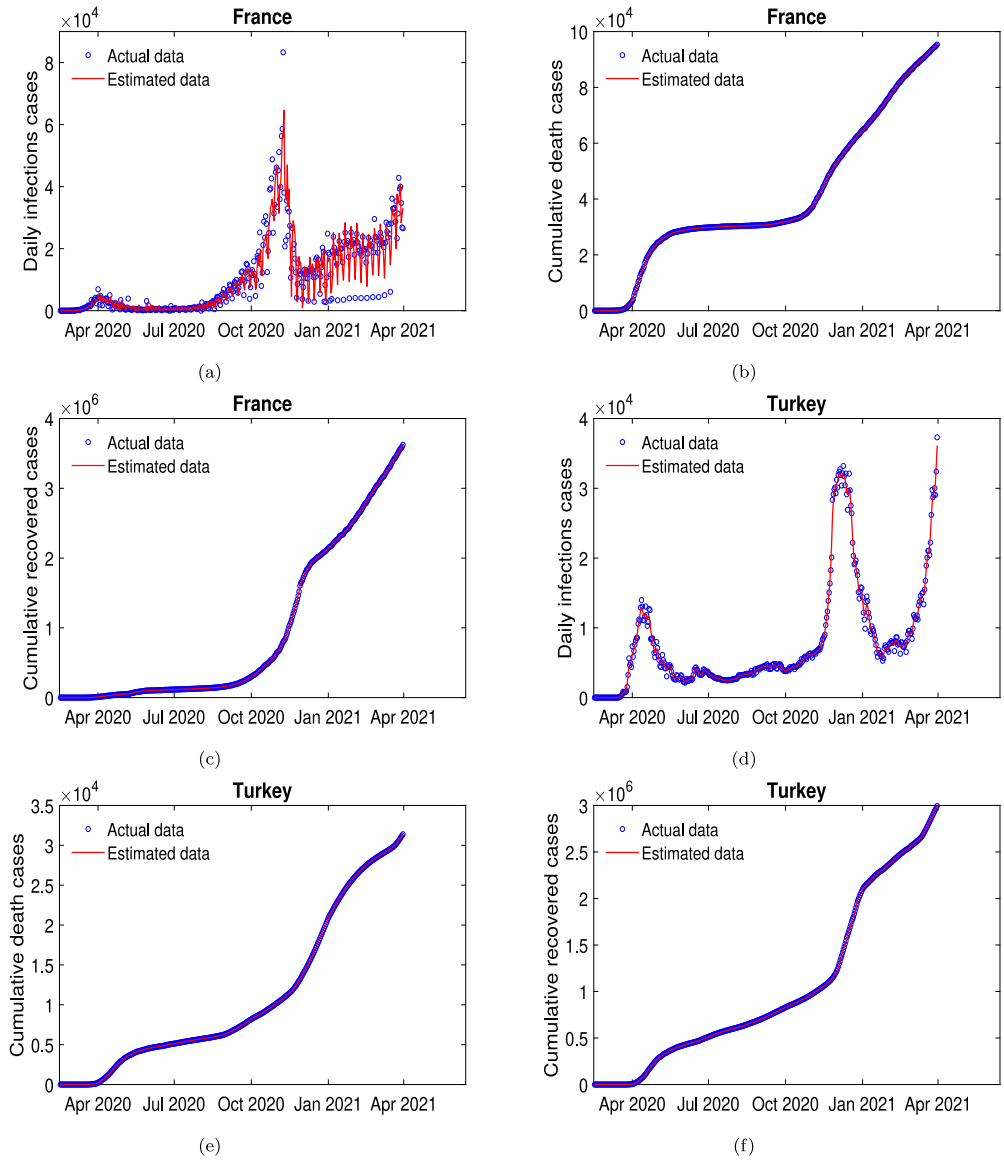


Fig. 8. Actual COVID-19 data and the estimated values as obtained using the proposed modeling scheme for 8(a) daily infections, 8(b) death cases, and 8(c) recovered cases for France, 8(d) daily infections, 8(e) death cases, and 8(f) recovered cases for Turkey.

Susceptible–Infected–Recovered for Asymptomatic–Symptomatic and Dead (SIRASD) model is used for Brazil COVID-19 data of 25th February 2020 to 30th March 2020 considering the long and short term effects of social distancing. In [33], the best data-fitted curves have been obtained using the Gaussian mixture model (GMM) and composite logistic growth model (CLGM) to find the optimum SIRD model for COVID-19. SIRD model parameters are derived as time-varying quantities, which is closer to the real-life scenario and can capture the inherent changes in the characteristic of the pandemic with time. The above changes can be due to various government policies, restrictions on domestic and international travel, quarantine rules imposed and also due to the medical facilities available in every country. The number of Gaussian (or LGM) waves and the parameters for each wave is estimated by the minimization of the objective function given by the sum of squares for residuals of values [7,43]. The minimization process uses the simplex search method in order to estimate the optimal values of the unknown parameters. Finally, the time-varying parameters of the SIRD model are computed from (1) as

$$\beta = -\frac{\mathrm{d}S}{\mathrm{d}t}\cdot\frac{N}{IS}, \qquad \gamma = \frac{\mathrm{d}R}{\mathrm{d}t}\cdot\frac{1}{I}, \tag{2a}$$

$$\mu = \frac{\mathrm{d}D}{\mathrm{d}t}\cdot\frac{1}{I}, \qquad \text{and} \quad \mathfrak{R}_0 = \frac{\beta S}{N(\gamma+\mu)}. \tag{2b}$$

2.2. ARIMA modeling

ARIMA model is a very popular time-domain model and has been used by various authors to model the prevalence or incidence of diseases such as SARS, HAV, Malaria, HFRS, Tuberculosis, Pertussis, Hepatitis, SFTS, HBV, Influenza, Human Brucellosis, Infectious Diarrhea, and Dengue Fever [44–49]. It has been used by several authors to predict the cumulative COVID-19 infections, the number of deaths reported, and the recovered cases for different countries.

In [29], ARIMA model is used to forecast the prevalence and incidence of COVID-19 for the next two days using the data from 20th January 2020 to 10th February 2020. Results were presented with 95% confidence. Data in the considered time range did not

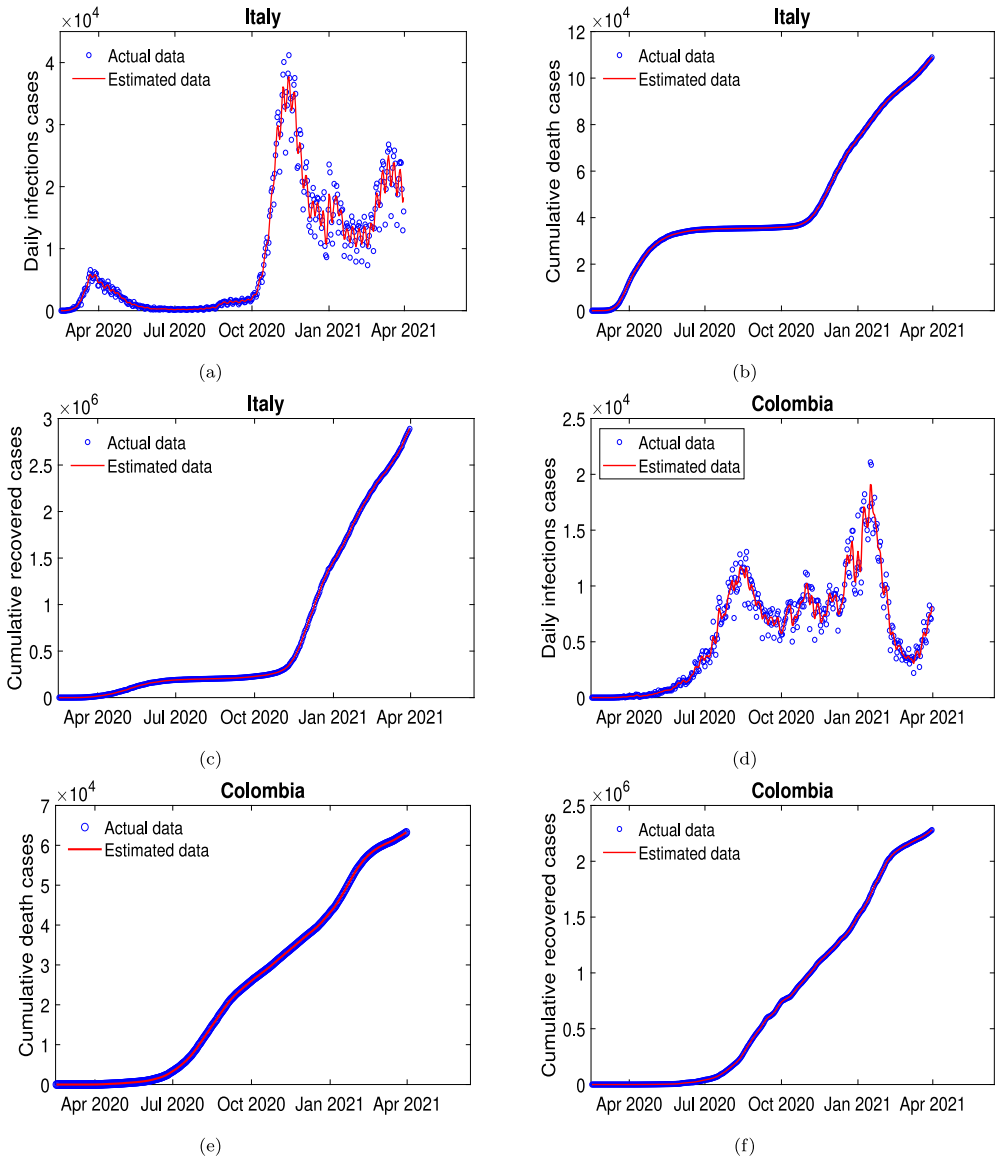


Fig. 9. Actual COVID-19 data and the estimated values as obtained using the proposed modeling scheme for 9(a) daily infections, 9(b) death cases, and 9(c) recovered cases for Italy, 9(d) daily infections, 9(e) death cases, and 9(f) recovered cases for Colombia.

present any seasonality. ARIMA (1,0,4) and ARIMA (1,0,3) models were selected as the best fit models. In [30], ARIMA model is used to capture the daily confirmed cases in Italy from 20th February, 2020 to 4th April, 2020. The seasonality of the data was tested using the Augmented Dickey–Fuller (ADF) test and the modified ADF-GLS (or ERS) test for unit root. The order of ARIMA was determined using Akaike's information criterion (AIC) and the mean absolute error (MAE). Perone further performed diagnostic tests on the residual data obtained using the selected ARIMA model including the Doornik and Hansen test for normality, Engle's Lagrange Multiplier test for the ARCH (autoregressive conditional heteroskedasticity) effect, and the Ljung–Box test for the autocorrelation.

The ARIMA model has been used in [35] to forecast the next ten days' cases using the data available from 31st January 2020 to 25th March 2020. Also, a nonlinear autoregressive neural network was used to forecast the next 50 days' data. Bayesian Information Criteria (BIC) was used to select ARIMA (1,1,0) model. It was mentioned that the autocorrelation function (ACF), and the partial autocorrelation function (PACF) can be used to choose the best fit and autocorrelation can be used to perform a diagnostic test on the residual signal. It was also mentioned in this work that BCI criteria is another method employed for model selection. The authors predicted that the number of new infections by 24th May 2020 will reach 1500. However, the actual numbers reached were 7113. In [36], data of 15 countries was considered from 21st January 2020 to 24th April 2020. The countries included were: United States, United Kingdom, Turkey, China, Russia, Netherlands, Switzerland, Germany, Iran, Brazil, Spain, Italy, France, Canada, and Belgium. Confirmed cases, recovered cases, and the deaths reported were modeled using the ARIMA model. Authors estimated that by 7th July 2020, the confirmed cases, deaths, and recoveries would be doubled in all countries considered in the study except China, Switzerland, and Germany. For the United States, the cumulative confirmed cases on 7th July 2020 were 3.33 times the cases on 24th April 2020, for the United Kingdom the data became 2.21 times, Turkey 1.98 times, Brazil 31.60 times, Spain 1.14 times, Russia 10.12 times, France 1.37 times, Italy 1.25 times, Canada 2.42 times, and Belgium 1.34 times.

In [28], ARIMA(2,1,1) model was used to obtain a four-week prediction for per day new infections in Saudi Arabia. It was estimated that the per day cases will reach 7,668 by 21*st* May

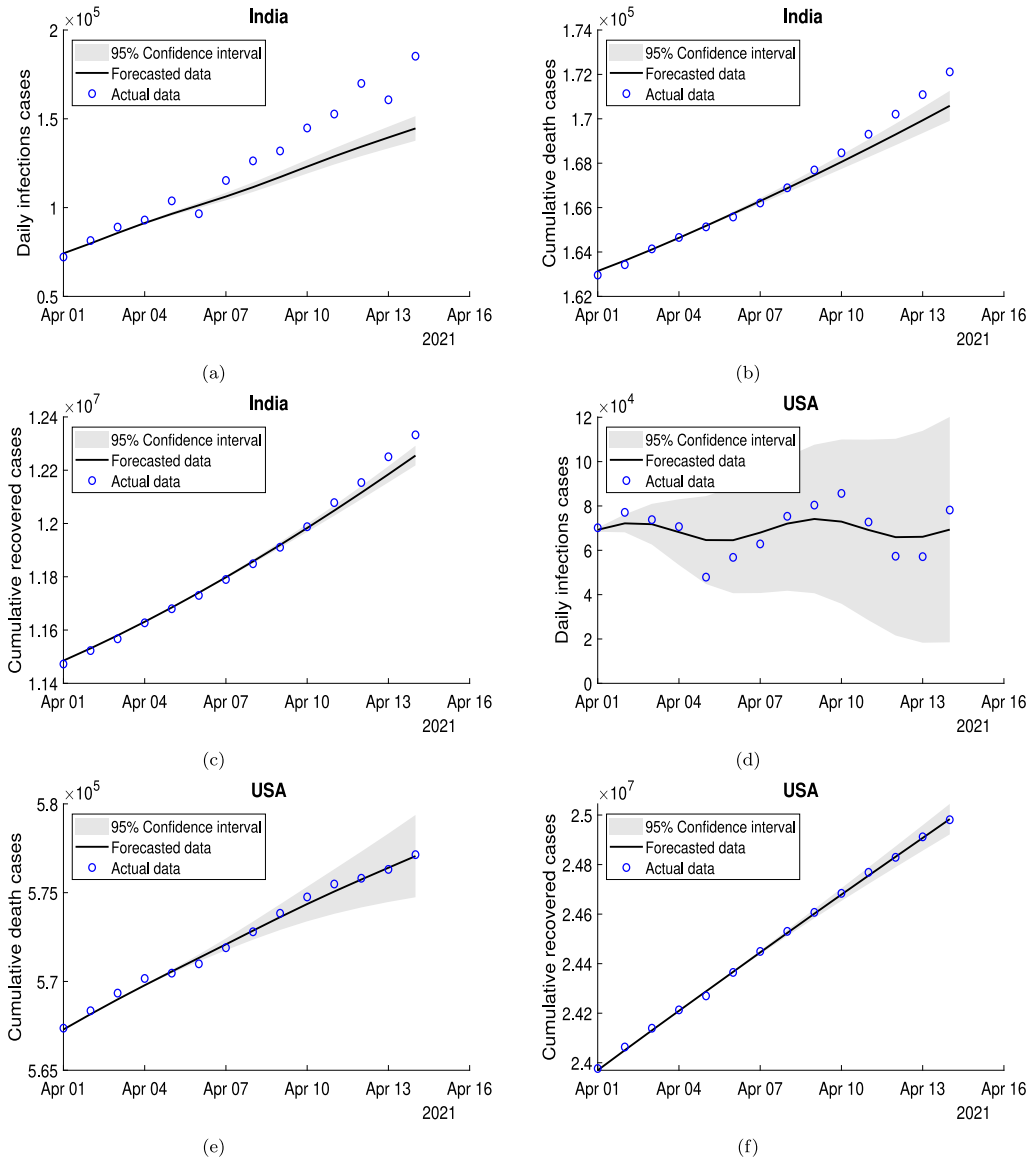


Fig. 10. Actual COVID-19 data and the forecasted values as obtained using the proposed modeling scheme for 10(a) daily infections, 10(b) death cases, and 10(c) recovered cases for India, 10(d) daily infections, 10(e) death cases, and 10(f) recovered cases for USA.

2020. However, the cases reported were 2,532. In [37], ARIMA (0,2,1), ARIMA (1,2,0), and ARIMA (0,2,1) were used to model prevalence of COVID-19 in Italy, Spain, and France, respectively. The models were selected based on the lowest MAPE values. Data from 21st February 2020 to 15th April 2020 was used and the total confirmed cases for the next ten days were predicted. The data of France was predicted with an RMSE of 9.1762e3, Italy was predicted with 2.1004e3 RMSE and the data of Spain was predicted with an RMSE of 3.2774e4. In [38], the spatial distribution of COVID-19 in Indian districts were analyzed and the prevalence and incidence of the disease were predicted using the ARIMA(2,2,2) model. Data from 30th January 2020 to 26th April 2020 was used to predict the data from 27th April 2020 to 11th May 2020.

In [9], a relationship between the number of COVID-19 cases and the population of the country is illustrated. Data of 145 countries have been modeled and the countries are grouped based on their proximity to each other. The study assumes that the spread of the disease is affected by various measurable and non-measurable factors that will remain similar in countries closer to each other. The average RMSE obtained in this case was 144.8. In [10], outbreaks of COVID-19 in Japan and South Korea were modeled using ARIMA(6,1,7) and ARIMA(2,1,3), respectively, for the duration from 20th January 2020 to 26th April 2020. The number of new infections per day for the next seven days was forecasted.

2.3. Multi-wave curve fitting model

In general, the evolution of the reported cases is modeled as a single wave (single peak wave). However, fitting the data with only one wave may not always be correct, since there are usually several waves with multiple peaks of the epidemic, while one wave captures very less fluctuations present in the data [7,8,14,19].

Some recent works are based on the assumption that multiple waves of a different peak, amplitude, and shape emerge and vanish overtime during the epidemic duration [7,8,14]. These works decompose the evolution of reported cases into several basic 'waves', where each basic wave is considered as a representation of the epidemic, both localized in time and position. Every single wave is considered as one of the known growth models such as

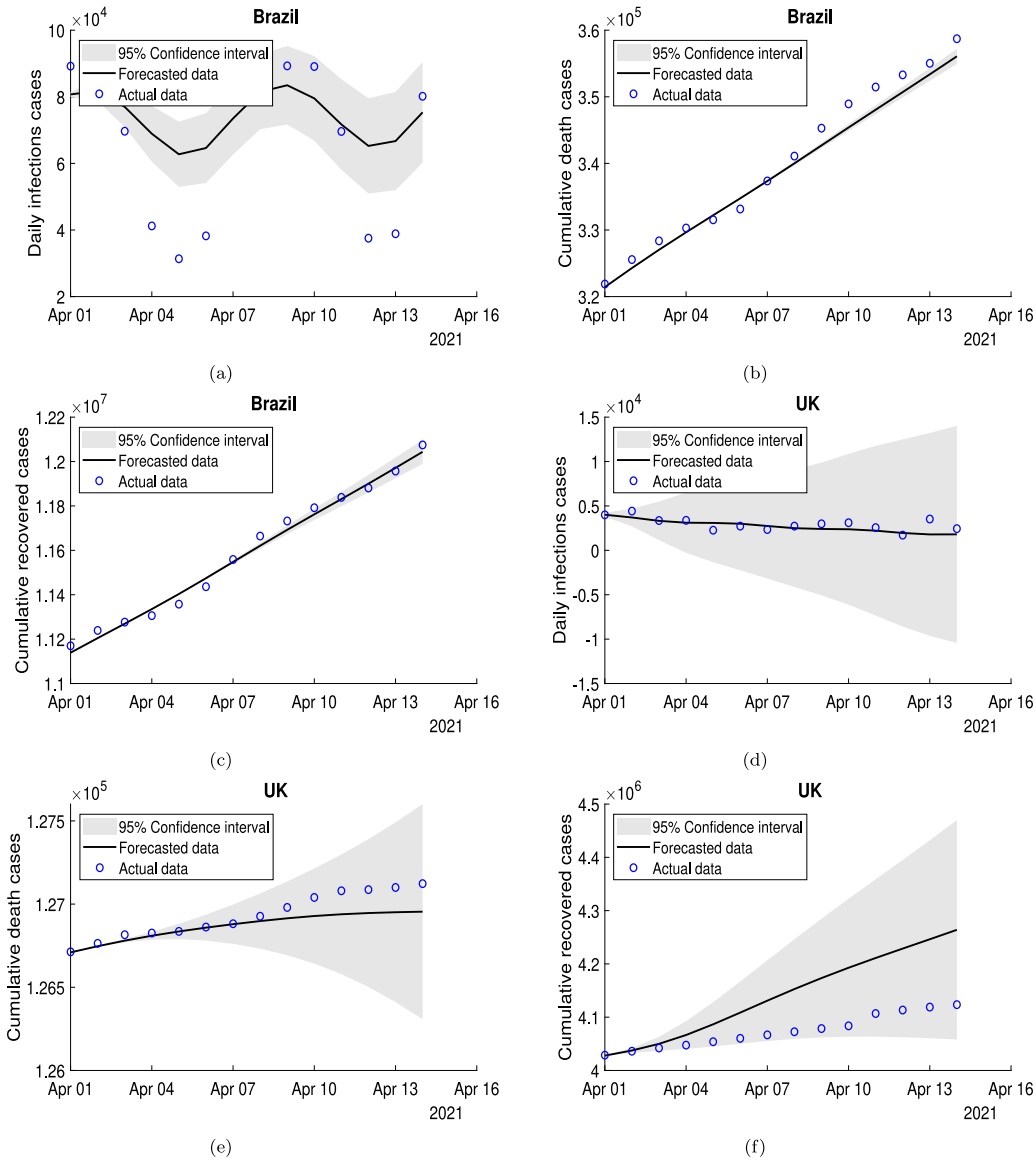


Fig. 11. Actual COVID-19 data and the forecasted values as obtained using the proposed modeling scheme for 11(a) daily infections, 11(b) death cases, and 11(c) recovered cases for Brazil, 11(d) daily infections, 11(e) death cases, and 11(f) recovered cases for UK.

the logistic or Gaussian. Briefly, we explain below both of these models.

2.3.1. Composite logistic growth model

First, we present the modeling framework with a single wave, i.e., with $P = 1$ for the logistic growth model. The logistic growth model is often used in epidemiology to model the spread of the infection [7,50,51]. Here, the number of infections initially grows exponentially, but later declines as the numbers approach the population's carrying capacity, where the carrying capacity is denoted as the number of people that can be infected eventually in a population. The cumulative number of infections on the tth day, denoted as $C(t)$, using the logistic growth model can be written as

$$C(t) = \frac{K}{1 + Ae^{-rt}}, \tag{3}$$

where K is the carrying capacity, A denotes the number of persons initially infected, and r is the growth rate. Corresponding to this model, the number of infected persons on the tth day, $I(t)$, is given by

$$I(t) = \frac{\mathrm{d}C(t)}{\mathrm{d}t} = \frac{KAre^{-rt}}{(1 + Ae^{-rt})^2}. \tag{4}$$

For any country, the numbers reported on day-0 (day of reference) are those that are active on that day. Hence, these are the cumulative numbers until that day and are equal to $C(0)$. Substituting $t = 0$ in (3) and (4), we obtain $C(0) = C_0 = \frac{K}{1+A}$ that implies $A = \frac{K}{C_0} - 1$, while $C_\infty = K$. Also, $C(0) - \epsilon = I(0)$. $I(0) = \frac{KAr}{(1+A)^2}$, and $A = \frac{K}{C_0} - 1$. The values of K, A, r, C_0 and I_0 are determined from the curve fitting of the available data.

The composite logistic growth model can be written as [52]

$$C(t) = \sum_{i=1}^{P} \frac{K_i}{1 + A_i e^{-r_i(t-\tau_i)}}, \tag{5}$$

where the number of waves P, and the four parameters (K_i, A_i, r_i, τ_i) for each wave are estimated by minimization of the objective function, which is the sum of squares of residuals [7, 43,53]. The minimization uses the simplex search method [54] to estimate optimal values of these unknown model parameters.

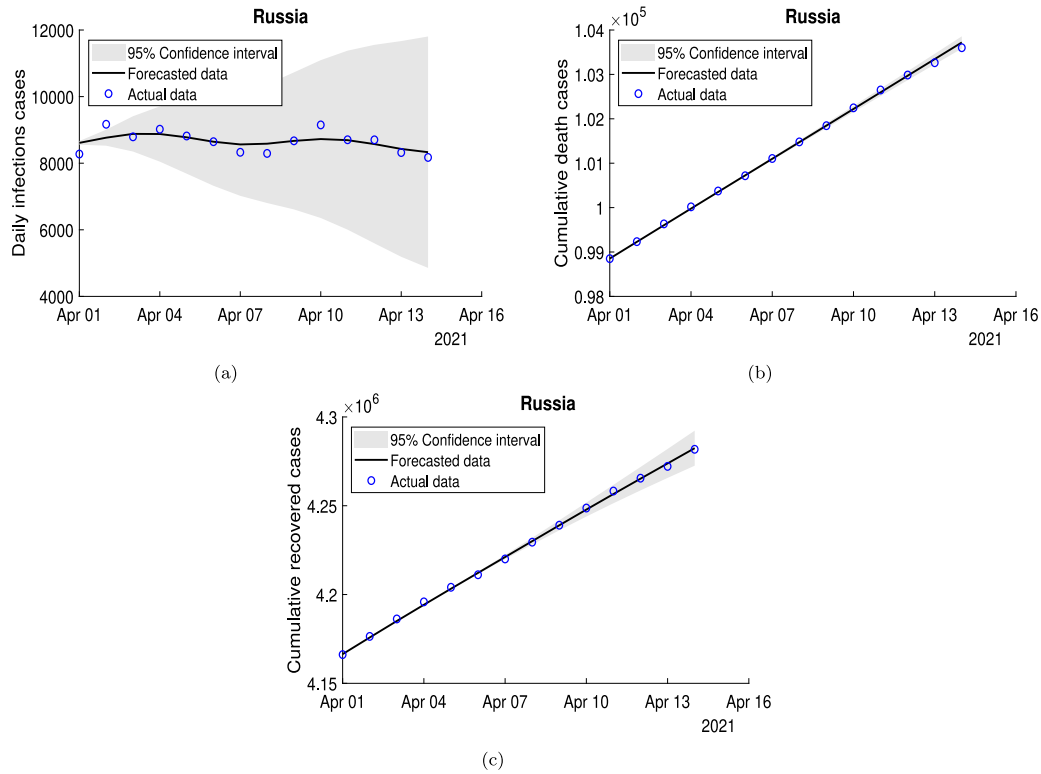


Fig. 12. Actual COVID-19 data and the forecasted values as obtained using the proposed modeling scheme for 10(a) daily infections, 10(b) death cases, and 10(c) recovered cases for India, 10(d) daily infections, 10(e) death cases, and 10(f) recovered cases for USA, 11(a) daily infections, 11(b) death cases, and 11(c) recovered cases for Brazil, 11(d) daily infections, 11(e) death cases, and 11(f) recovered cases for UK, 12(a) daily infections, 12(b) death cases, and 12(c) recovered cases for Russia.

Table 1
Summary of state-of-the-arts techniques used for COVID-19 data modeling.

Methodology	Ref.	Strength	Weakness
Compartmental models (SIR, SIRD, etc.)	[14–16,32–34]	1. Model the spread of the disease using the interaction between different population compartments which is a more natural model. 2. Provide estimates of the important parameters like infection rate, recovery rate, death rate, and reproduction number of the epidemic. 3. Reproduction number is the most useful for planning and deciding control strategies.	These models make two assumptions: (i) the chance of any infected person to infect other susceptible persons is constant during the epidemic duration, and (ii) assume that every infected person has a constant chance to recover at any given time. Both of these may not be true. Moreover, a precise and closed-form solution of all the system parameters is difficult to obtain.
ARIMA	[9,10,28–30,35–38]	It is a parametric model which can fit non-stationary data.	It does not perform well if the correlation in data samples is negligible.
Multiwave Fitting (Gaussian and Logistic)	[7,8,14].	These are parametric models which can capture multiple emerging waves of epidemic.	Require predefined shape of the waves and estimations of parameters for fitting.
Dictionary Learning	[23]	It is a non-parametric model which can capture any shape of epidemic wave.	Requires large amounts of data. The training process is expensive involving selection of large number of hyperparameters.

2.3.2. Composite Gaussian model

Next, we model $I(t)$ using the Gaussian function. Here, the number of infected persons $I(t)$ on the tth day is given by

$$I(t) = \alpha\, e^{-\frac{(t-\mu)^2}{2\sigma^2}}, \tag{6}$$

where μ denotes the mean, σ^2 denotes the variance of the Gaussian function, while $I_0 = \alpha\, e^{-\frac{\mu^2}{2\sigma^2}}$. Thus, the composite Gaussian model can be written as

$$I(t) = \sum_{i=1}^{P} I_i(t) = \sum_{i=1}^{P} \alpha_i \exp\left(-\frac{(t-\mu_i)^2}{2\sigma_i^2}\right), \tag{7}$$

where regression parameters α_i, μ_i and σ_i are the amplitude, mean, and standard deviation, respectively.

The model is utilized for a maximum of five epidemic waves. The sum of all the waves should predict the main reported cases.

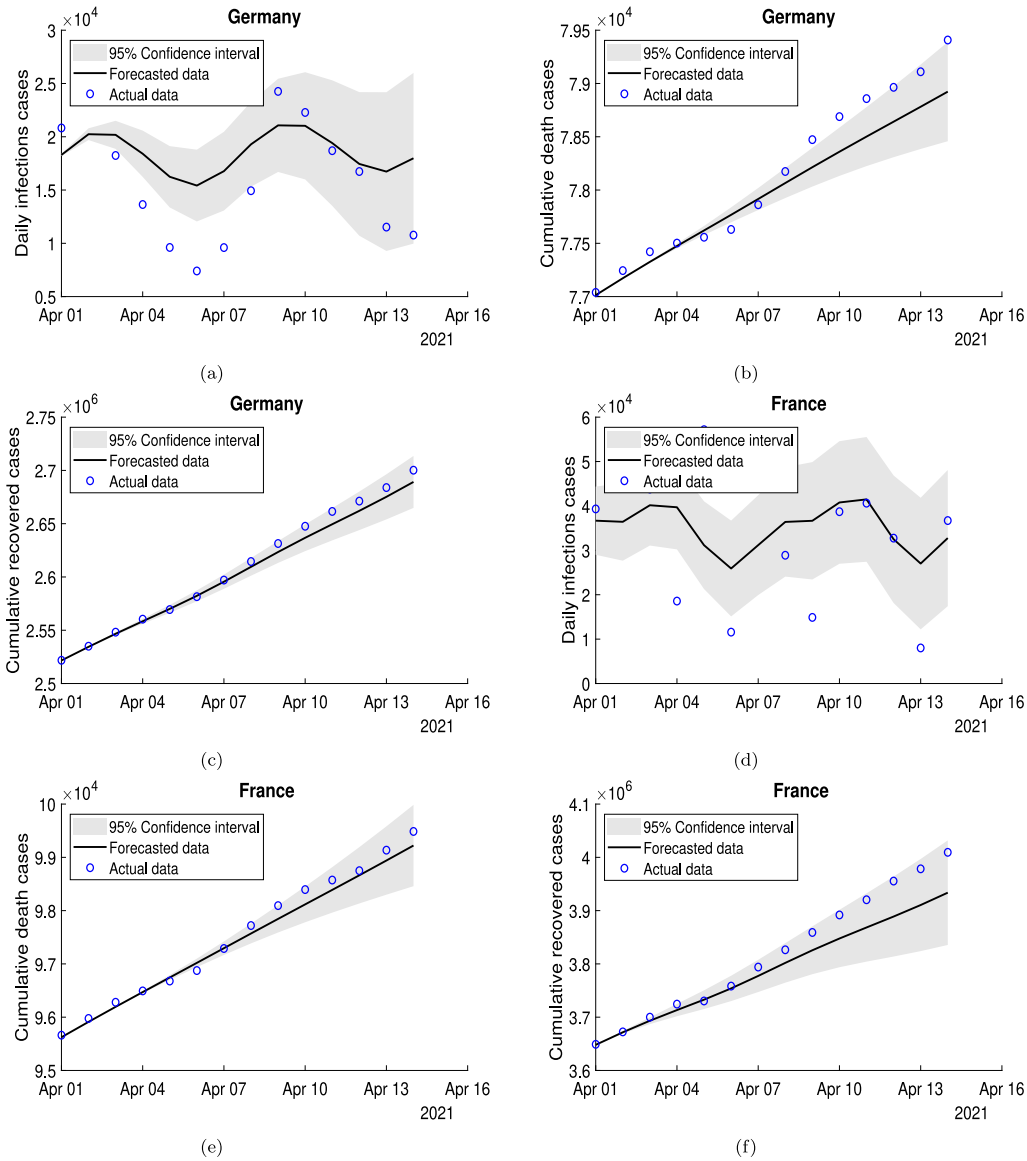


Fig. 13. Actual COVID-19 data and the forecasted values as obtained using the proposed modeling scheme for 13(a) daily infections, 13(b) death cases, and 13(c) recovered cases for Germany, 13(d) daily infections, 13(e) death cases, and 13(f) recovered cases for France.

2.4. Dictionary learning model

In one of the recent works [23], forecasting of COVID-19 is done using dictionary learning and ONMF. This approach mainly consists of four steps. First, the dictionary is learned by minibatch learning from the entire duration of COVID-19 data, followed by, progressively adapting and improving the learned dictionary via ONMF. Later, a one-step prediction is made by partially fitting a learned dictionary to the known data so as to get a forecasted value of one day ahead. Lastly, by recursively applying the one-step predictions, extrapolation of predictions for the near future is done. We name this method as ONMF, from here onwards.

The pseudo-code of the method is as follows:

- *Dictionary learning*: The first step deals with dictionary learning. Consider the number of days T for which the data $\mathbf{x} \equiv (x_1, x_2, \ldots, x_T)$ is available. Random patches of length N are extracted from this data and stacked as columns of matrix $\mathbf{X}$.

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d] \in \mathbb{R}^{N\times d}, \tag{8}$$

where $\mathbf{x}_i \in \mathbb{R}^{N\times 1}$. Given a data matrix $\mathbf{X}$, the goal is to find nonnegative dictionary $\mathbf{W} \in \mathbb{R}^{N\times r}$ and nonnegative code matrix $\mathbf{H} \in \mathbb{R}^{r\times d}$ by solving the following optimization problem:

$$\inf_{\mathbf{W}\in\mathbb{R}^{N\times r},\mathbf{H}\in\mathbb{R}^{r\times d}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \lambda \|\mathbf{H}\|_1 , \tag{9}$$

where $\|\mathbf{A}\|_F^2 = \sum_{i,j} A_{ij}^2$ denotes the matrix Frobenius norm and $\lambda \geq 0$ is the regularization parameter. $\mathbf{W}$ represents the learned dictionary having r number of atoms and $\mathbf{H}$ represents the code matrix. Above optimization problem is also known as the Nonnegative matrix factorization (NMF) problem.

- *Refining the learned dictionary*: In the second stage, the learned dictionary is further updated using online NMF. Here, Online implies learning the sequence of dictionary matrices from the sequence of data matrices $\mathbf{X}$, generated by moving one day ahead and considering all previous data points.
- *Forecasting*: Further, a learned dictionary is used to predict one day ahead data by partial fitting and updating

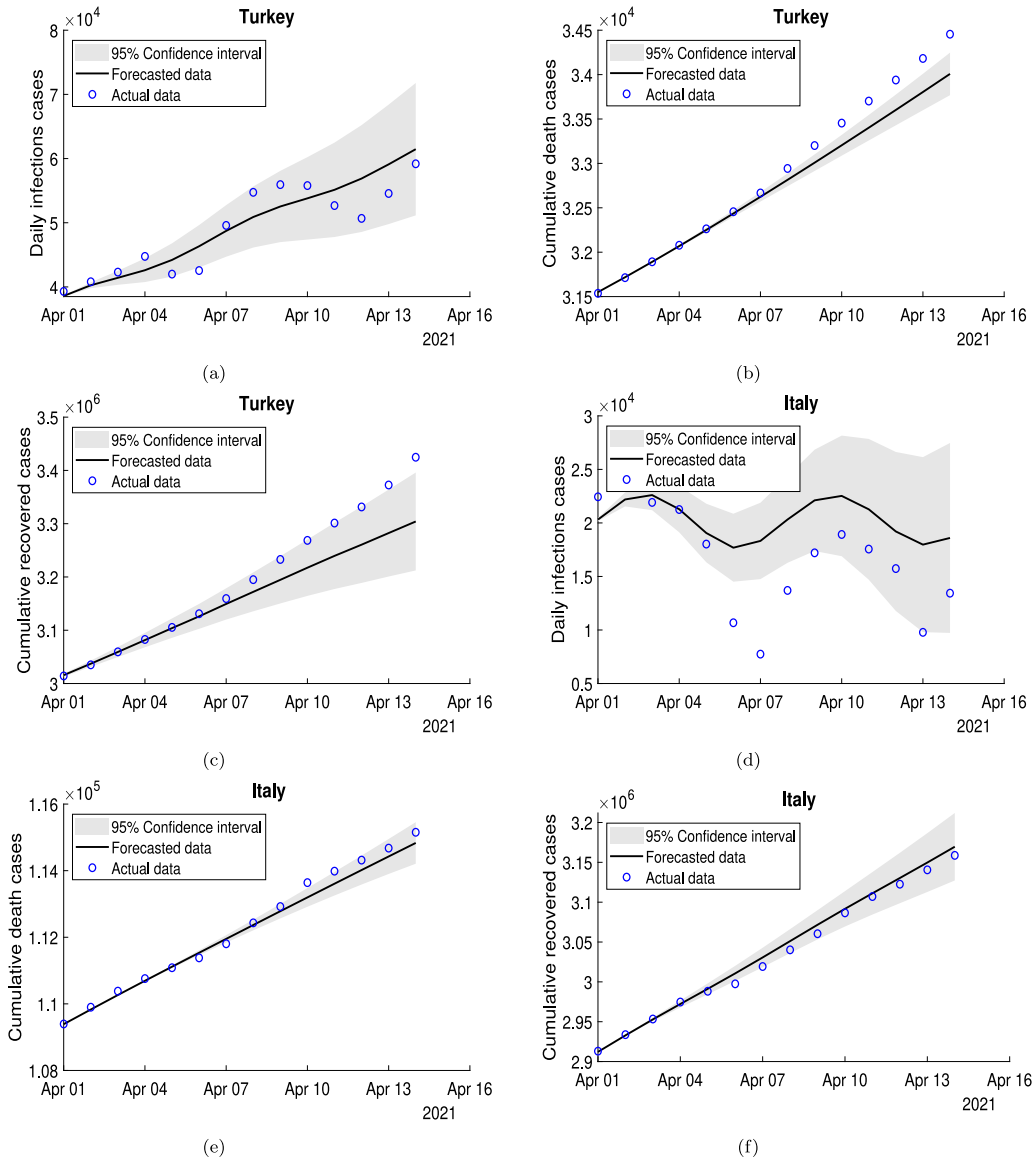


Fig. 14. Actual COVID-19 data and the forecasted values as obtained using the proposed modeling scheme for 13(a) daily infections, 13(b) death cases, and 13(c) recovered cases for Germany, 13(d) daily infections, 13(e) death cases, and 13(f) recovered cases for France, 14(a) daily infections, 14(b) death cases, and 14(c) recovered cases for Turkey, 14(d) daily infections, 14(e) death cases, and 14(f) recovered cases for Italy.

H. For more details of this framework, please refer to the work by [23]. By recursively using the one-step predictions, extrapolation of the future values is carried out.

2.5. Strengths and weaknesses of forecasting models

To summarize, several approaches have been proposed by researchers to predict the COVID-19 outbreak including SIR models, and variants, ARIMA modeling, multi-wave curve fitting modeling and dictionary learning modeling. Among all, compartment models (SIR and its variants) are the most frequently used approach so far [55] and dictionary learning modeling is the least used method in the literature for COVID-19 forecasting. In essence, all these methods exhibit many pros and cons, which are described in Table 1.

3. Proposed model

In this work, we propose modeling based on the ARIMA model to forecast the daily infections, cumulative deaths, and cumulative recovered cases of COVID-19. Since the COVID-19 data for various countries is non-stationary, traditional ARMA methods may not be sufficient to capture the data efficiently. In such cases, ARIMA performs better by capturing the trend or seasonality in the data. To further improve the performance of ARIMA models, we propose to add a pre-processing step to the ARIMA model and estimate the trend in the data using a low pass Gaussian filter as shown in Fig. 1.

In [27], a method is proposed to estimate an ARIMA model. ARIMA models are considered as a generalization of ARMA models to predict a given time series data using its past values. While ARMA models are used to fit the time-series data with stationary property, ARIMA has been developed for data with inherent non-stationarity or with seasonality trends. The ARIMA model can be understood in two steps, where the first step removes the non-stationary trend from the data, and the second step models the output obtained from the first step using an ARMA model. The model is denoted as ARIMA(p, d, q), where p denotes the order of AR, q is the order of MA, and d represents the degree of difference used in the model, mathematically, expressed as follows:

$$y(t) = (1 - B)^d x(t), \tag{10}$$

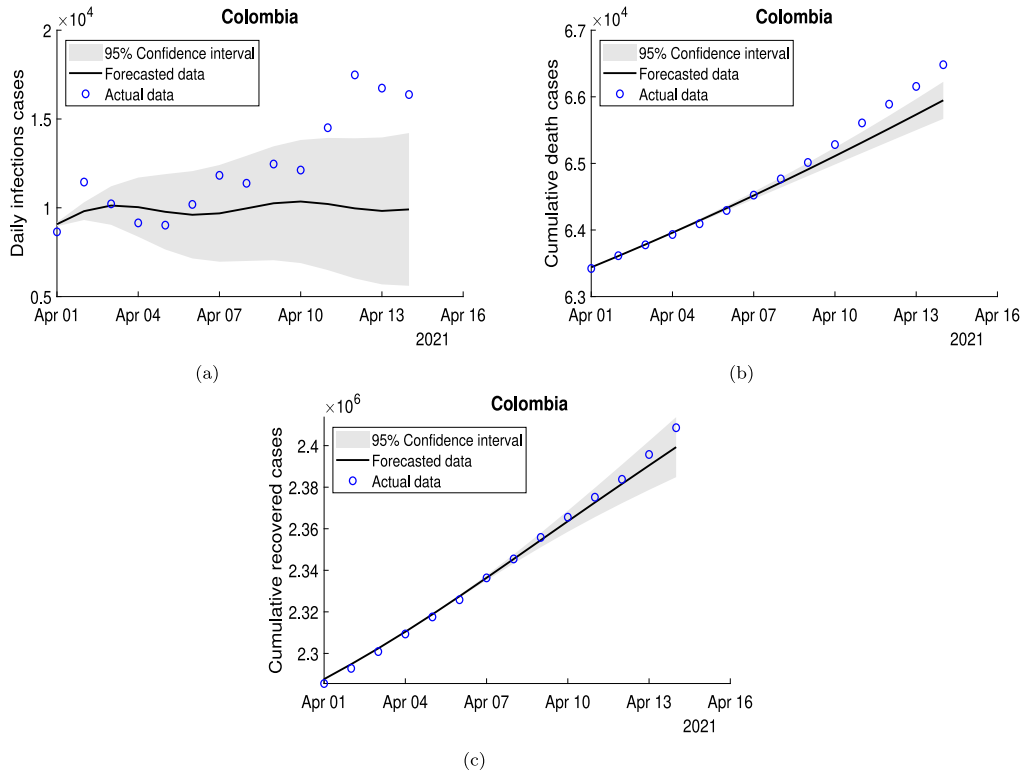


Fig. 15. Actual COVID-19 data and the forecasted values as obtained using the proposed modeling scheme for 15(a) daily infections, 15(b) death cases, and 15(c) recovered cases for Colombia.

$$\left(1-\sum_{k=1}^{p} a(k)B^k\right) y(t) = \left(1+\sum_{k=1}^{q} b(k)B^k\right) e(t), \tag{11}$$

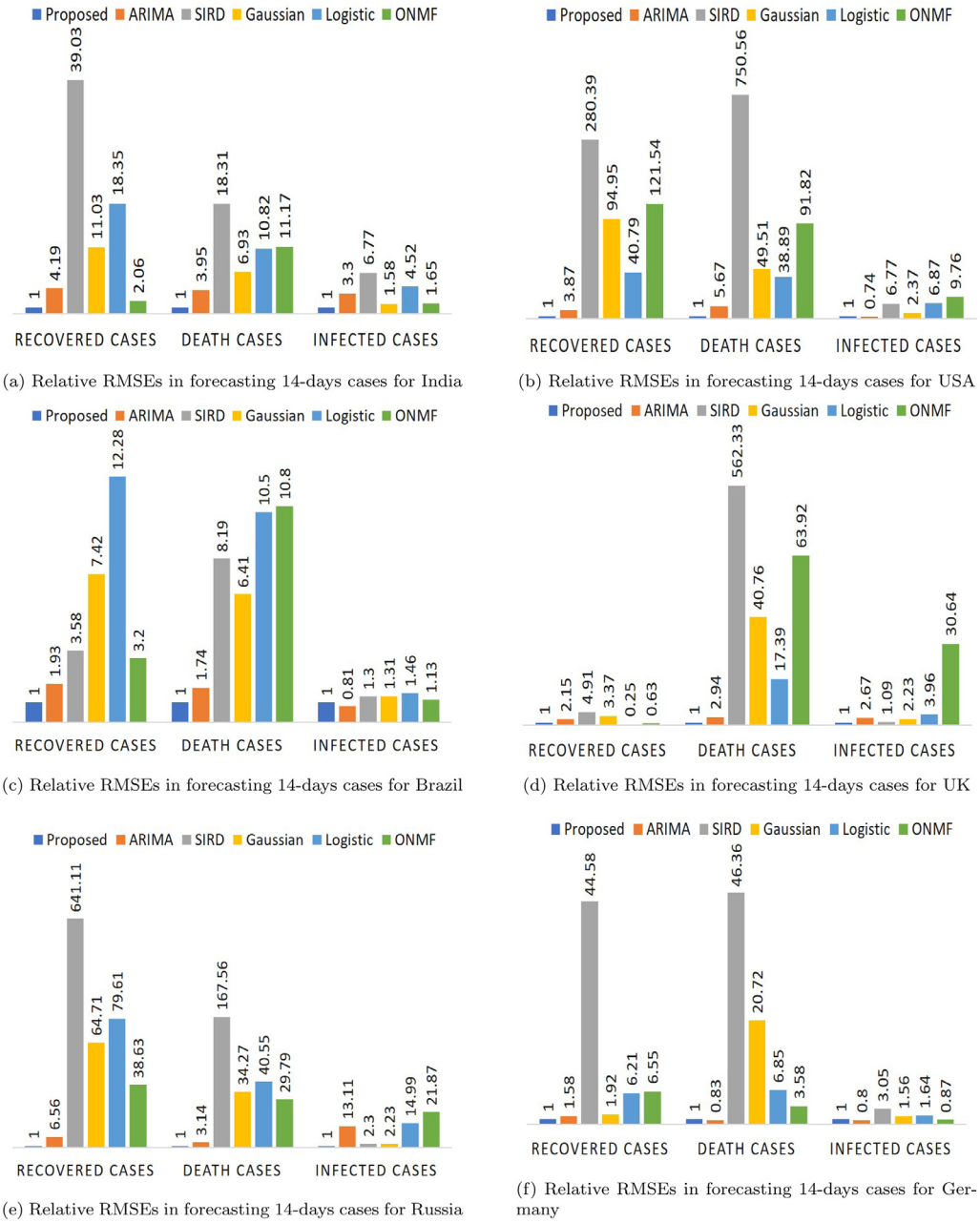
where B is the back-shift operator, B^p represents back-shift by p-steps, and $x(t)$ is the non-stationary time series data to be modeled. In the first step, (10), signal $x(t)$ is converted to a stationary signal $y(t)$. Step 2 models the signal $y(t)$ using an ARMA model as depicted in (11), where $a(k)$ are the auto-regressive model coefficients, $b(k)$ are the moving average coefficients, and $e(t)$ is the error signal.

The Box–Jenkins method used for the modeling of data with the best ARIMA model involves three steps: model selection, parameter estimation, and model validation. For a given time-series, the order of the model can be estimated using the sample autocorrelation function (ACF) and partial autocorrelation function (PACF). If the data has an inherent seasonal trend, the initial differencing step is used once or more than once to convert the data to a stationary series. The output of the differencing step is checked after every iteration using the ACF plot. For a stationary series, the values of the ACF should rapidly converge to zero. If it is not so, the differencing step should be repeated. The number of times the differencing step is repeated to obtain the stationarity provides the order value for d. ACF is also used to estimate the MA order, q, with the assumption that for pure MA processes ACF values converge to zeros after lag q. Similarly, it can be observed that for a pth order pure AR processes values of the PCF become zero after lag p. Akaike information criterion (AIC) or the Bayesian Information Criterion (BIC) can be used to obtain the optimum order of ARMA, where the objective is to minimize AIC or BIC values. Once the values for p, q, and d are obtained, the model can be estimated using either the maximum likelihood estimation or the least-squares estimation.

In this work, we first filter the given time-series using a Gaussian filter to obtain its low pass version which is then modeled using ARIMA. Here, the stationarity of the low pass signal is checked, and accordingly the parameter value for d is obtained. The stationary signal obtained after differencing is estimated using ARMA and optimum p and q parameters are selected. The cut-off frequency of the low pass filter is changed and the best ARIMA models are obtained for each case. It is pertinent to mention that since the ARIMA model is developed for the low pass version, whereas our main goal is to estimate the given time-series and forecast the future values, metrics proposed in the Box–Jenkins ARIMA modeling method such as BIC may not be a correct choice. BIC and AIC values will prefer the model that can estimate the low pass signal efficiently, but not the original series. For this reason, we have used several different metrics such as RMSE, MAPE (mean absolute percentage error), MAE (mean absolute error), MARE (mean absolute relative error), RMSRE (root mean square relative error), MSPE (mean square percentage error), and MSE (mean square error) to select the optimum model. We can empirically select the metric that gives the model which estimates the original series the best. If the obtained ARIMA model can estimate the time-series efficiently, the residual signal should be white and this fact can also be used to validate the model. Statistical tests on the residual value can be used such as Ljung–Box Q test, Box–Pierce test, Breusch–Godfrey test, and the Durbin–Watson test. In this work, we have used the Ljung–Box Q test for checking the estimated model.

4. Results

In this section, we present the simulation results for modeling and predicting the COVID-19 data for ten countries, including India, the USA, the UK, Russia, Brazil, Germany, France, Turkey, Italy, and Colombia. The data used in this paper involves the cumulative recovered cases, cumulative death cases, and number of new infections as collected from the Worldometer [56] and WHO daily situation report [57]. The data from February 15 2020 till April 14 2021 has been used for modeling and the data for the



(a) Relative RMSEs in forecasting 14-days cases for India

(b) Relative RMSEs in forecasting 14-days cases for USA

(c) Relative RMSEs in forecasting 14-days cases for Brazil

(d) Relative RMSEs in forecasting 14-days cases for UK

(e) Relative RMSEs in forecasting 14-days cases for Russia

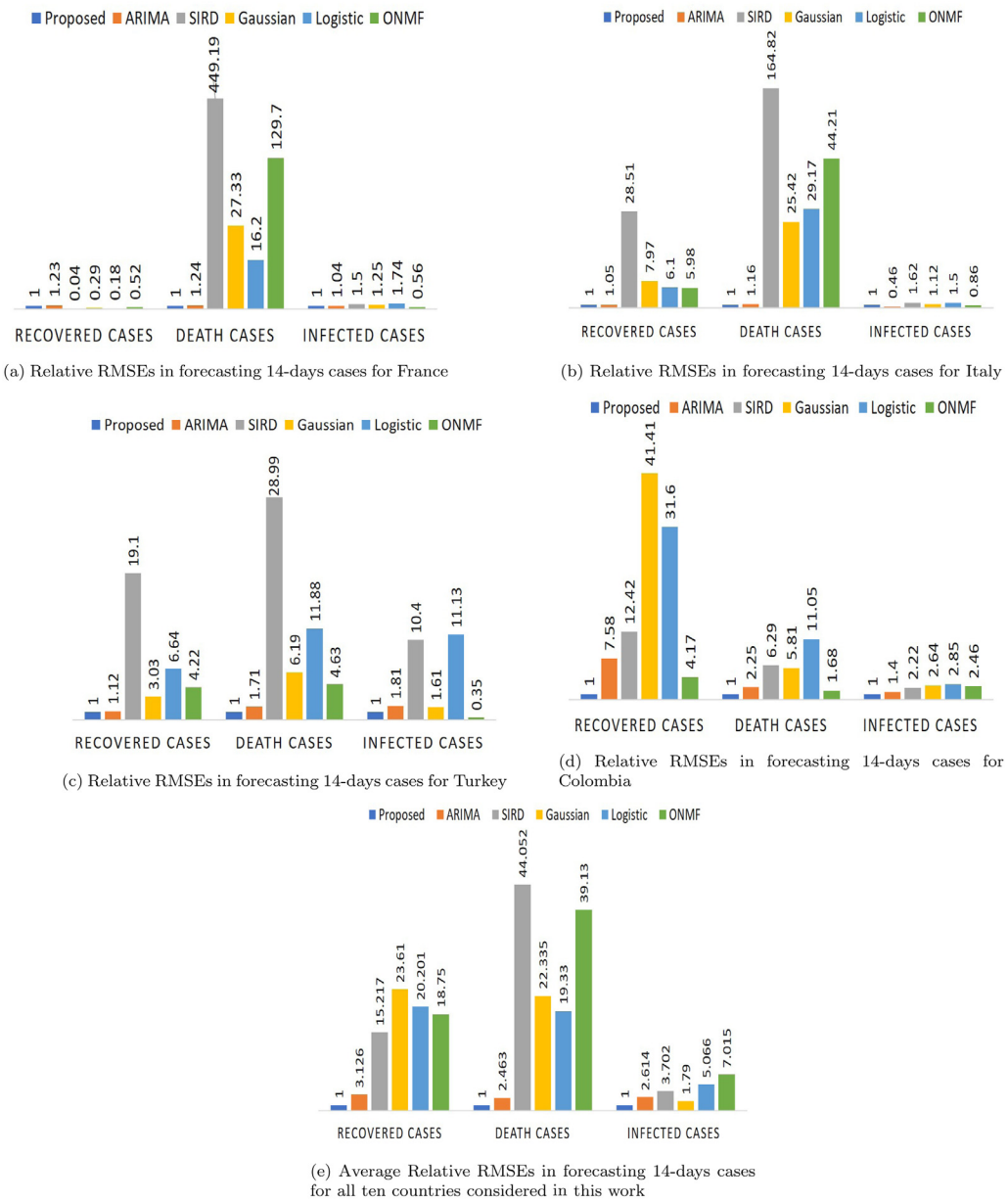
(f) Relative RMSEs in forecasting 14-days cases for Germany

Fig. 16. A comparison of the proposed model in terms of relative RMSEs in forecasting next 14-days cases (recovered, death and infected cases) with ARIMA, SIRD, Gaussian growth, composite Logistic growth, and ONMF models for 16(a) India, 16(b) USA, 16(c) Brazil, 16(d) UK, 16(e) Russia, and 16(f) Germany.

next 14 days is used for validating the prediction performance of the model. We compare the performance of the proposed method with the ARIMA model, SIRD model, composite Gaussian model, composite logistic growth model, and ONMF model. These models are used to forecast the next 7 days and 14 days data and the RMSE values obtained are compared in Table 2. Results for the SIRD model have been obtained using the model discussed by [33] and results for the ARIMA model are obtained using the Box–Jenkins approach [27].

Figs. 2 and 3 show the normalized RMSE values obtained when different metrics are used to select the optimum ARIMA model to estimate the cumulative recovered data and death data for the ten countries considered. Using this, we observed that models selected on the basis of BIC value and RMSRE values provide the best estimates in most of the cases for cumulative recovered and death cases, respectively, and thus, these metrics were selected for the given data series. Furthermore, Fig. 4 shows normalized RMSE values obtained when different metrics are used to select the optimum ARIMA model to estimate the daily infection data. Here, it is observed that the models selected using the MSPE values are optimum in most of the cases and thus, MSPE is the chosen metric for the daily infection data.

The proposed method has been used to model the COVID-19 data from February 15, 2020 till March 31, 2021. Figs. 5–7 shows the actual data, estimated data and the forecasted data for the next 14 days as obtained using the proposed model for India, USA, Brazil, UK and Russia, Germany and Figs. 8–9 includes the graphs for France, Turkey, Italy and Colombia. Figs. 12–15 show the performance of the proposed algorithm in forecasting the short-term future data of the ten countries considered. The figures



(a) Relative RMSEs in forecasting 14-days cases for France

(b) Relative RMSEs in forecasting 14-days cases for Italy

(c) Relative RMSEs in forecasting 14-days cases for Turkey

(d) Relative RMSEs in forecasting 14-days cases for Colombia

(e) Average Relative RMSEs in forecasting 14-days cases for all ten countries considered in this work

Fig. 17. A comparison of the proposed model in terms of relative RMSEs in forecasting next 14-days cases (recovered, death and infected cases) with ARIMA, SIRD, Gaussian growth, composite Logistic growth, and ONMF models for 17(a) France, 17(b) Italy, 17(c) Turkey, 17(d) Colombia, and 17(e) an average of the ten countries.

include the actual data, the output of the Gaussian filter, and the forecasted data with 95% confidence interval. Table 3 provides the order of the ARIMA models obtained with the proposed method.

The relative RMSE of a model is defined here as the ratio of the RMSE obtained using the model to RMSE obtained with the proposed method. This gives a metric to compare the RMSEs obtained using different models and compare the performance of the proposed method as compared to existing modeling techniques. The relative RMSEs so obtained are plotted as histograms in Figs. 16–17 for all ten countries and also the average obtained after removing outliers is plotted as well.

5. Discussion

Compared to the existing methodologies, the models obtained using the proposed methodology predict the future cases far better for India, UK, Russia, and Colombia as shown in Table 2 and Figs. 16(a), 16(d), 16(e), 17(d). For India, the performance of ARIMA for both 7 days and 14 days forecasts is very poor compared to the proposed methodology. As compared to the RMSE obtained using the proposed method, RMSE obtained with ARIMA is 4.19 times for recovered cases, 3.95 times for death cases, and 3.3 times for infected cases. In these cases, we observe that the proposed method estimates the low pass version of the actual data efficiently, as shown in Figs. 5–9. Also, the future cases follow this low-pass trajectory as shown in Fig. 12 and Fig. 15. The reported cases depend on various external factors such as socio-economic activities, policy changes, festivals, holidays, local weather, etc., and also on the number of testing being done. These changes may cause sudden fluctuations in the time series. In the case of India, the future cases were not dependent on these fluctuations and therefore, estimation using a low pass trend of the time-series produced better prediction results. In cases of number of new infections, the performance of Gaussian curve fitting is closer to the proposed method with its 1.58 times RMSE as compared to the proposed method.

Table 2
RMSE values obtained for predicting the 7 days and 14 days ahead cases using the proposed methodology for ten most affected countries.

Country	Model	RMSE					
		Recovered		Death		Infected	
		7 days	14 days	7 days	14 days	7 days	14 days
India	Proposed	**9.25e3**	**3.10e4**	**120.11**	**611.73**	**5.05e3**	**1.89e4**
	ARIMA	3.97e4	1.30e5	901.55	2.41e3	3.01e4	6.24e4
	SIRD	9.63e5	1.21e6	8.47e3	1.12e4	9.38e4	1.28e5
	Gaussian	1.55e5	3.42e5	1.97e3	4.24e3	1.28e4	2.99e4
	Logistic	3.33e5	5.69e5	4.02e3	6.62e3	5.20e4	8.54e4
	ONMF	2.84e4	6.39e4	4.13e3	6.83e3	1.97e4	3.12e4
USA	Proposed	**9.73e3**	**8.31e3**	**259.37**	**250.48**	7.57e3	7.83e3
	ARIMA	1.25e4	3.22e4	640.03	1.42e3	**3.39e3**	**5.81e3**
	SIRD	2.34e6	2.33e6	1.88e5	1.88e5	4.65e4	5.30e4
	Gaussian	3.56e5	7.89e5	7.89e3	1.24e4	1.47e4	1.86e4
	Logistic	2.23e5	3.39e5	7.89e3	9.74e3	4.74e4	5.38e4
	ONMF	9.37e5	1.01e6	8.72e3	2.30e4	4.78e4	7.64e4
Brazil	Proposed	3.11e4	**3.02e4**	**1.00e3**	**1.99e3**	1.97e4	1.79e4
	ARIMA	**1.87e4**	5.83e4	1.74e3	2.89e3	**1.85e4**	**1.45e4**
	SIRD	1.50e5	1.08e5	8.19e3	1.37e4	2.40e4	2.32e4
	Gaussian	1.49e5	2.24e5	6.41e3	1.28e4	2.30e4	2.34e4
	Logistic	2.38e5	3.71e5	1.05e4	1.43e4	2.33e4	2.62e4
	ONMF	5.36e4	9.68e4	1.08e4	1.81e4	1.12e4	2.02e4
UK	Proposed	**3.37e4**	**8.25e4**	**16.35**	**88.56**	**461.83**	**656.06**
	ARIMA	9.23e4	1.78e5	108.49	260.69	1.32e3	1.75e3
	SIRD	1.08e5	4.05e5	1.37e4	4.98e4	549.65	717.55
	Gaussian	1.02e5	2.78e5	2.29e3	3.61e3	1.39e3	1.46e3
	Logistic	7.81e3	2.08e4	1.47e3	1.54e3	2.72e3	2.60e3
	ONMF	6.02e4	5.19e4	2.15e3	5.66e3	1.54e4	2.01e4
Russia	Proposed	**1.03e3**	**1.02e3**	**21.02**	**44.64**	**288.94**	**293.62**
	ARIMA	2.54e4	6.69e3	44.66	140.42	288.94	293.63
	SIRD	6.48e5	6.54e5	7.48e3	7.53e3	3.40e3	3.85e3
	Gaussian	4.09e4	6.60e4	1.53e3	3.29e3	549.68	674.91
	Logistic	6.08e4	8.12e4	1.81e3	2.36e3	4.11e3	4.40e3
	ONMF	3.63e4	3.94e4	1.33e3	1.57e3	4.33e3	6.42e3
Germany	Proposed	**1.20e3**	**6.76e3**	76.91	239.44	5.41e3	4.75e3
	ARIMA	3.53e3	1.06e4	**36.11**	**199.61**	**4.48e3**	**3.78e3**
	SIRD	2.56e5	3.00e5	1.06e4	1.11e4	1.32e4	1.45e4
	Gaussian	8.49e3	1.29e4	1.88e3	4.96e3	7.64e3	7.44e3
	Logistic	2.23e4	4.18e4	1.08e3	1.644e3	6.90e3	7.81e3
	ONMF	2.52e4	4.41e4	867.09	857.56	3.07e3	4.15e3
France	Proposed	**8.22e3**	3.93e4	74.75	**158.06**	**1.68e4**	1.44e4
	ARIMA	1.24e4	4.83`e4	**66.39**	195.75	**1.68e4**	1.50e4
	SIRD	1.52e3	**1.70e3**	7.00e4	7.10e4	2.63e4	2.16e4
	Gaussian	6.63e3	1.12e4	1.88e3	4.32e3	2e4	1.80e4
	Logistic	4.80e3	6.90e3	1.79e3	2.56e3	3.02e4	2.5e4
	ONMF	1.22e4	2.05e4	1.22e4	2.05e4	5.14e3	**8.08e3**
Turkey	Proposed	**4.42e3**	**5.09e4**	**21.85**	**218.01**	1.93e3	3.01e3
	ARIMA	8.42e3	5.72e4	77.92	372.51	2.70e3	5.46e3
	SIRD	9.62e5	9.72e5	6.11e3	6.32e3	2.52e4	3.13e4
	Gaussian	7.66e4	1.54e5	734.78	1.35e3	3.91e3	4.85e3
	Logistic	2.35e5	3.38e5	1.75e3	2.59e3	2.61e4	3.35e4
	ONMF	1.37e5	2.15e5	502.25	1.01e3	**828.96**	**1.05e3**
Italy	Proposed	6.75e3	7.88e3	99.29	**221.45**	4.91e3	5.14e3
	ARIMA	**4.55e3**	**7.54e3**	**77.11**	256.32	**2.94e3**	**2.36e3**
	SIRD	1.69e5	2.15e5	3.52e4	3.65e4	9.62e3	8.33e3
	Gaussian	2.85e4	6.01e4	2.46e3	5.63e3	5.54e3	5.77e3
	Logistic	3.45e4	4.60e4	5.02e3	6.46e3	8.58e3	7.73e3
	ONMF	2.87e4	3.98e4	5.75e3	9.79e3	2.79e3	4.41e3
Colombia	Proposed	**1.53e3**	**3.26e3**	**25.80**	**228.91**	**1.11e3**	**3.62e3**
	ARIMA	9.20e4	2.47e4	115.71	513.95	2.38e3	5.08e3
	SIRD	5.28e4	4.05e4	1.68e3	1.4438e3	5.17e3	8.02e3
	Gaussian	8.44e4	1.35e5	525.69	1.33e3	6.83e3	9.56e3
	Logistic	6.81e4	1.03e5	1.65e3	2.53e3	7.66e3	1.03e4
	ONMF	1.01e4	1.36e4	445.71	384.33	6.27e3	8.91e3

The proposed method predicts the recovered cases for Germany with the least RMSE and provides considerable improvement over ARIMA. For daily infections and death cases, ARIMA works better, as shown in Fig. 16(f). However, the RMSE values are very close, 0.8 times of the proposed method for infected cases and 0.83 times for death cases. For France, the proposed algorithm works better than ARIMA for recovered cases and equivalently for daily infected cases. However, for recovered cases SIRD gives the best performance with 0.04 relative RMSE and ONMF gives 0.56 relative RMSE for infected cases. For the COVID-19 cases of the USA, the proposed methodology forecasts the recovered cases and death cases far better than the existing methods by a factor of 10^1 and 10^3, respectively. However, for daily infected cases, ARIMA estimates the future cases better than the proposed method by a very small margin with relative RMSE of 0.74. This is also observed for Brazil daily infected cases, where the performance of ARIMA is slightly better than the proposed method, with 0.81 relative RMSE. As observed from Fig. 10(d) and Fig. 11(a), the actual cases have fluctuations and thus, the low pass model obtained using the proposed method is not able to predict these high-frequency changes. For Turkey, the performance of the proposed prediction algorithm is superior for cumulative recovered cases and death cases. For daily infection cases, it performs better than ARIMA but falls short when compared to ONMF. The ARIMA model predicts the future cases for Italy with the least RMSE. However, the performance is very close to that of the proposed method.

Considering an average over all ten countries, Fig. 17(e), the proposed method predicts the recovered cases with 0.32 times RMSE as compared to ARIMA, 0.07 times of SIRD, 0.04 times that of composite Gaussian growth model, 0.05 times composite Logistic growth model, and ONMF model. For prediction of death cases, the proposed method predicts the 14 days data with RMSE 0.40 times as compared to ARIMA, 0.02 times that of SIRD, 0.04 times composite Gaussian growth model, 0.05 times composite Logistic growth model, and 0.03 times of RMSE obtained with ONMF model. The performance of the proposed method in predicting daily infected cases is compared using RMSE values, which is 0.38 times that of RMSE obtained when the ARIMA model is used, 0.27 times that of SIRD, 0.56 times of composite Gaussian growth model, 0.2 times composite Logistic growth model, and 0.14 times the RMSE of ONMF model.

6. Conclusion and future work

In this work, we reviewed and benchmarked the most popular modeling techniques of COVID-19 data estimation and continuous prediction. These models are ARIMA, SIRD, composite Gaussian growth model, composite Logistic growth model, and dictionary learning model (i.e., ONMF). Composite Gaussian and Logistic methods model the COVID-19 data by a number of overlapping Gaussian and Logistic distribution waves, where each basic wave is localized in time and considered as a representation of the epidemic. However, the assumption of having similar waves throughout the epidemic duration may not give realistic forecasts. Therefore, to overcome this drawback, we also reviewed the recently proposed model-free approach of dictionary learning. This method learns the waves directly from the data without assuming any predefined shape. We also proposed a new data modeling strategy by estimating a trend of the data and then using an optimized ARIMA model. The trend is obtained using a low-pass Gaussian filter. The performance of these models was compared based on the RMSE values obtained for the 7-days and 14-days ahead prediction, and it was shown that for most of the cases the performance of the proposed methodology is far superior to the other existing methodologies. The number of daily COVID-19 infections, the cumulative number of recovered cases, and the cumulative deaths reported for India, the USA, the UK, Russia, Brazil, Germany, France, Italy, Turkey, and Colombia

Table 3
ARIMA model as used in the proposed methodology to develop models for different countries for cumulative recovered cases, cumulative death cases, and daily infected cases.

Country	Recovered cases		Death cases		Daily infections	
	Model	RMSRE	Model	BIC	Model	MSPE
India	ARIMA(6,1,3)	0.06	ARIMA(8,2,7)	6.01	ARIMA(4,1,3)	3.16e3
USA	ARIMA(9,2,9)	0.23	ARIMA(9,1,8)	20.17	ARIMA(7,2,7)	2.66e3
Brazil	ARIMA(9,2,8)	0.26	ARIMA(6,2,9)	108.46	ARIMA(6,2,7)	3.65e4
Russia	ARIMA(7,2,8)	0.07	ARIMA(1,1,10)	263.24	ARIMA(8,2,3)	503.46
UK	ARIMA(9,2,3)	0.30	ARIMA(4,1,7)	56.44	ARIMA(7,2,8)	301.72
Germany	ARIMA(7,1,2)	0.15	ARIMA(3,2,6)	6.04	ARIMA(9,2,4)	9.83e5
France	ARIMA(5,2,3)	0.85	ARIMA(9,2,10)	6.80	ARIMA(10,2,7)	3.28e7
Turkey	ARIMA(10,2,9)	8.31	ARIMA(8,2,1)	9.00	ARIMA(6,2,4)	1.82e3
Italy	ARIMA(8,2,2)	0.25	ARIMA(3,2,6)	9.97	ARIMA(9,2,2)	1.27e3
Colombia	ARIMA(9,2,8)	0.09	ARIMA(2,1,5)	6.58	ARIMA(2,2,8)	725.13

have been considered for modeling and continuous prediction. Although we have considered these mentioned countries only in our study, the proposed methodology can be used for the continuous monitoring and prediction of COVID-19 of any country, state, and region.

So far, a number of methods have been proposed for forecasting COVID-19 data. The utility of these methods is still limited for long-term forecasting and predicting onsets of COVID-19 waves, and in exploring the data correlation geographically. Furthermore, it is also imperative to provide insight into any model with respect to assessment, planning, and policy-making for combating the spread of COVID-19. Policy data integration with forecasting models is another important work that is worth exploring in the future as policies and public health guidelines issued at the state and local level could aid in the advancement of forecasting models and increasing accuracy of forecasting as well as expanding the potential impact of the forecasts on policy decisions. Real-time live forecasting is one of the primary research areas to explore in the future. Hybrid algorithms based on the proposed model can be explored for the same.

CRediT authorship contribution statement

Binish Fatimah: Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review & editing. **Priya Aggarwal:** Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review & editing. **Pushpendra Singh:** Conceptualization, Methodology, Visualization, Validation, Writing - review & editing. **Anubha Gupta:** Conceptualization, Methodology, Visualization, Validation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Authors also would like to thanks anonymous reviewers and editors for providing constrictive review comments to improve the manuscript.

References

[1] H. Zhu, L. Wei, P. Niu, The novel coronavirus outbreak in Wuhan, China, Glob. Health Res. Policy 5 (1) (2020) 6, http://dx.doi.org/10.1186/s41256-020-00135-6.

[2] A. Bhaik, V. Singh, E. Gandotra, D. Gupta, Detection of improperly worn face masks using deep learning - A preventive measure against the spread of COVID-19, Int. J. Interact. Multimed. Artif. Intell. InPress (2021) 1–12, http://dx.doi.org/10.9781/ijimai.2021.09.003.

[3] B.H. Meyer, B. Prescott, X.S. Sheng, The impact of the COVID-19 pandemic on business expectations, Int. J. Forecast. (2021) http://dx.doi.org/10.1016/j.ijforecast.2021.02.009.

[4] J.P. Ioannidis, S. Cripps, M.A. Tanner, Forecasting for COVID-19 has failed, Int. J. Forecast. (2020) http://dx.doi.org/10.1016/j.ijforecast.2020.08.004.

[5] M. Dur-e-Ahmad, M. Imran, Transmission dynamics model of coronavirus COVID-19 for the outbreak in most affected countries of the world, Int. J. Interact. Multimed. Artif. Intell. InPress (2020) 1, http://dx.doi.org/10.9781/ijimai.2020.04.001.

[6] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong, J. He, Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions, J. Thoracic Disease 12 (3) (2020).

[7] M. Batista, Estimation of a state of corona 19 epidemic in august 2020 by multistage logistic model: A case of EU, USA, and world (update september 2020), MedRxiv (2020) arXiv:https://www.medrxiv.org/content/early/2020/10/06/2020.08.31.20185165.full.pdf.

[8] A. Singhal, P. Singh, B. Lall, S.D. Joshi, Modeling and prediction of COVID-19 pandemic using Gaussian mixture model, Chaos Solitons Fractals 138 (2020) 110023.

[9] A. Hernandez-Matamoros, H. Fujita, T. Hayashi, H. Perez-Meana, Forecasting of COVID19 per regions using ARIMA models and polynomial functions, Appl. Soft Comput. 96 (2020) 106610.

[10] X. Duan, X. Zhang, ARIMA modelling and forecasting of irregularly patterned COVID-19 outbreaks using Japanese and South Korean data, Data Brief 31 (2020) 105779.

[11] S. Shastri, K. Singh, S. Kumar, P. Kaur, V. Mansotra, Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study, Chaos Solitons Fractals 140 (2020) 110227, http://dx.doi.org/10.1016/j.chaos.2020.110227.

[12] H. Abbasimehr, R. Paki, Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization, Chaos Solitons Fractals 142 (2021) 110511, http://dx.doi.org/10.1016/j.chaos.2020.110511.

[13] F. Brauer, Compartmental models in epidemiology, in: F. Brauer, P. van den Driessche, J. Wu (Eds.), Mathematical Epidemiology, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 19–79, http://dx.doi.org/10.1007/978-3-540-78911-6_2.

[14] P. Singh, A. Gupta, Generalized SIR (GSIR) epidemic model: An improved framework for the predictive monitoring of COVID-19 pandemic, ISA Trans. (2021) S0019–0578(21)00099–9, http://dx.doi.org/10.1016/j.isatra.2021.02.016, 33610314[pmid].

[15] D. Sen, D. Sen, Use of a modified SIRD model to analyze COVID-19 data, Ind. Eng. Chem. Res. 60 (11) (2021) 4251–4260, http://dx.doi.org/10.1021/acs.iecr.0c04754, http://arxiv.org/abs/DOI:10.1021/acs.iecr.0c04754.

[16] G.C. Calafiore, C. Novara, C. Possieri, A time-varying SIRD model for the COVID-19 contagion in Italy, Annu. Rev. Control 50 (2020) 361–372.

[17] S. Annas, M. Isbar Pratama, M. Rifandi, W. Sanusi, S. Side, Stability analysis and numerical simulation of SEIR model for pandemic COVID-19 spread in Indonesia, Chaos Solitons Fractals 139 (2020) 110072.

[18] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, T. Yang, B. Lou, Y. Chi, H. Long, M. Ma, Q. Yuan, S. Zhang, D. Zhang, F. Ye, J. Xin, Predicting COVID-19 in China using hybrid AI model, IEEE Trans. Cybern. 50 (7) (2020) 2891–2904, http://dx.doi.org/10.1109/TCYB.2020.2990162.

[19] T. Tat Dat, P. Frédéric, N.T.T. Hang, M. Jules, N. Duc Thang, C. Piffault, R. Willy, F. Susely, H.V. Lê, W. Tuschmann, N. Tien Zung, Epidemic dynamics via wavelet theory and machine learning with applications to Covid-19, Biology 9 (12) (2020) http://dx.doi.org/10.3390/biology9120477.

[20] P. Wang, X. Zheng, J. Li, B. Zhu, Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics, Chaos Solitons Fractals 139 (2020) 110058, http://dx.doi.org/10.1016/j.chaos.2020.110058.

[21] J. Schuttler, R. Schlickeiser, F. Schlickeiser, M. Kröger, Covid-19 predictions using a Gauss model, based on data from april 2, Physics 2 (2) (2020) 197–212, http://dx.doi.org/10.3390/physics2020013.
[22] T. Tat Dat, P. Frédéric, N.T.T. Hang, M. Jules, N. Duc Thang, C. Piffault, R. Willy, F. Susely, H.V. Lê, W. Tuschmann, N. Tien Zung, Epidemic dynamics via wavelet theory and machine learning with applications to Covid-19, Biology 9 (12) (2020) http://dx.doi.org/10.3390/biology9120477.
[23] H. Lyu, C. Strohmeier, G. Menz, D. Needell, COVID-19 time-series prediction by joint dictionary learning and online NMF, ArXiv E-Prints (2020) arXiv:2004.09112, arXiv:2004.09112.
[24] Z. Yu, X. Zheng, F. Huang, W. Guo, L. Sun, Z. Yu, A framework based on sparse representation model for time series prediction in smart city, Front. Comput. Sci. 15 (1) (2021) 151305, http://dx.doi.org/10.1007/s11704-019-8395-7.
[25] R. Rosas-Romero, A. Díaz-Torres, G. Etcheverry, Forecasting of stock return prices with sparse representation of financial time series over redundant dictionaries, Expert Syst. Appl. 57 (2016) http://dx.doi.org/10.1016/j.eswa.2016.03.021.
[26] A. Helmi, M.W. Fakhr, A.F. Atiya, Multi-step ahead time series forecasting via sparse coding and dictionary based techniques, Appl. Soft Comput. 69 (2018) 464–474, http://dx.doi.org/10.1016/j.asoc.2018.04.017.
[27] G. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day, 1976.
[28] S.I. Alzahrani, I.A. Aljamaan, E.A. Al-Fakih, Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions, J. Infect. Public Health 13 (7) (2020) 914–919.
[29] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, M. Ciccozzi, Application of the ARIMA model on the COVID-2019 epidemic dataset, Data Brief 29 (2020) 105340.
[30] G. Perone, An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in Italy, 2020, arXiv:2004.00382.
[31] R.R. Sharma, M. Kumar, S. Maheshwari, K.P. Ray, EVDHM-ARIMA-based time series forecasting model and its application for COVID-19 cases, IEEE Trans. Instrum. Meas. 70 (2021) 1–10, http://dx.doi.org/10.1109/TIM.2020.3041833.
[32] D. Fanelli, F. Piazza, Analysis and forecast of COVID-19 spreading in China, Italy and France, Chaos Solitons Fractals 134 (2020) 109761.
[33] P. Singh, A. Singhal, B. Fatimah, A. Gupta, An improved data driven dynamic SIRD model for predictive monitoring of COVID-19, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 8158–8162.
[34] S. Bastos, D. Cajueiro, Modeling and forecasting the early evolution of the Covid-19 pandemic in Brazil, Sci. Rep. 10 (2020).
[35] F.M. Khan, R. Gupta, ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India, J. Saf. Sci. Resil. 1 (1) (2020) 12–18, http://dx.doi.org/10.1016/j.jnlssr.2020.06.007, URL https://www.sciencedirect.com/science/article/pii/S2666449620300074.
[36] R. Singh, M. Rani, A. Bhagavathula, R. Sah, A. Rodriguez-Morales, H. Kalita, C. Nanda, S. Sharma, Y. Sharma, A. Rabaan, J. Rahmani, P. Kumar, The prediction of COVID-19 pandemic for top-15 affected countries using advance ARIMA model, JMIR Public Health Surveillance 6 (2020).
[37] Z. Ceylan, Estimation of COVID-19 prevalence in Italy, Spain, and France, Sci. Total Environ. 729 (2020) 138817.
[38] S. Roy, G. Bhunia, P. Shit, Spatial prediction of COVID-19 epidemic using ARIMA techniques in India, Model. Earth Syst. Environ. 7 (2020) 1–7.
[39] H.W. Hethcote, The mathematics of infectious diseases, SIAM Rev. 42 (4) (2000) 599–653, http://dx.doi.org/10.1137/S0036144500371907.
[40] P. Heidrich, T. Goetz, Modelling Dengue with the SIR Model, Vol. 30, 2019, pp. 175–182, http://dx.doi.org/10.1007/978-3-030-27550-1_22.
[41] O. Koutou, B. Traoré, B. Sangaré, Mathematical model of malaria transmission dynamics with distributed delay and a wide class of nonlinear incidence rates, in: Y. Rogovchenko (Ed.), Cogent Math. Stat. 5 (1) (2018) 1564531.
[42] L. Zhong, L. Mu, J. Li, J. Wang, Z. Yin, D. Liu, Early prediction of the 2019 novel coronavirus outbreak in the mainland China based on simple mathematical model, IEEE Access 8 (2020) 51761–51769.
[43] M. Batista, Estimation of the final size of the COVID-19 epidemic, MedRxiv Preprint (2020) 01–11.
[44] P. Guan, D.-S. Huang, B.-S. Zhou, Forecasting model for the incidence of hepatitis a based on artificial neural network, World J. Gastroenterol. : WJG 10 (2005) 3579–3582.
[45] A. Earnest, M. Chen, D. Ng, Y. Leo, Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore, BMC Health Serv. Res. 5 (2005) 36.
[46] J. Gaudart, O. Toure, N. Dessay, A. Dicko, S. Ranque, L. Forest, J. Demongeot, O. Doumbo, Modelling malaria incidence with environmental dependency in a locality of sudanese savannah area, Mali, Malar. J. 8 (2009) 61, http://dx.doi.org/10.1186/1475-2875-8-61.
[47] Q.-Y. Liu, X. Liu, B. Jiang, W. Yang, Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model, BMC Infect. Dis. 11 (2011) 218.
[48] X. Zhang, Y. Liu, M. Yang, T. Zhang, A. Young, X. Li, Comparative study of four time series methods in forecasting typhoid fever incidence in China, PLoS One 8 (2013) e63116.
[49] H. Ren, J. Li, Z.-A. Yuan, J.-Y. Hu, Y. Yu, Y. Lu, The development of a combined mathematical model to forecast the incidence of Hepatitis E in Shanghai, China, BMC Infect. Dis. 13 (2013) 421.
[50] F.A. Cássaro, L.F. Pires, Can we predict the occurrence of COVID-19 cases? Considerations using a simple model of growth, Sci. Total Environ. 728 (2020) 138834, http://dx.doi.org/10.1016/j.scitotenv.2020.138834.
[51] S. Petropoulos, Forecasting the novel coronavirus COVID-19, PLOS ONE 15 (3) (2020) 1–8, http://dx.doi.org/10.1371/journal.pone.0231236.
[52] R. Pearl, L.J. Reed, On the rate of growth of the population of the United States since 1790 and its mathematical representation, Proc. Natl. Acad. Sci. USA 6 (6) (1920) 275–288.
[53] M. Batista, Fitvirusxx, 2020, https://www.mathworks.com/matlabcentral/fileexchange/{76956}-fitvirusxx, MATLAB Central File Exchange, Retrieved August 28.
[54] J.C. Lagarias, J.A. Reeds, M.H. Wright, P.E. Wright, Convergence properties of the nelder-mead simplex method in low dimensions, SIAM J. Optim. 9 (1) (1998) 112–147.
[55] I. Rahimi, F. Chen, A.H. Gandomi, A review on COVID-19 forecasting models, Neural Comput. Appl. (2021) http://dx.doi.org/10.1007/s00521-020-05626-8.
[56] Worldometers, Worldometers, 2020, Retrieved: Oct. 17, 2020.
[57] WHO, World health organization, "coronavirus disease (COVID-2019) situation reports", 2020, Retrieved: Oct. 17, 2020..