medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Modeling and Forecasting of Covid-19 Growth Curve in India

Vikas Kumar Sharma* and Unnati Nigam

Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi, India

*Corresponding Email: vikasstats@rediffmail.com

Abstract:

In this article, we analyze the growth pattern of Covid-19 pandemic in India from March 4th to May 15th using regression analysis (exponential and polynomial), auto-regressive integrated moving averages (ARIMA) model as well as exponential smoothing and Holt-Winters models. We found that the growth of Covid-19 cases follows a power regime of (t^2, t,..) after the exponential growth. We found the optimal change points from where the Covid-19 cases shift their course of growth from exponential to quadratic and then from quadratic to linear. We have also found the best fitted regression models using the various criteria such as significant p-values, coefficients of determination and ANOVA etc. Further, we search the best fitting ARIMA model for the data using the AIC (Akaike Information Criterion) and CAIC (Consistent Akaike Information Criterion) and provide the forecast of Covid-19 cases for future days. We also use usual exponential smoothing and Holt-Winters models for forecasting purpose. We further found that the ARIMA (2,2,0) model is the best-fitting model for Covid-19 cases in India.

Keywords: Covid-19, Regression analysis, Exponential growth, Polynomial growth, ANOVA, ARIMA, Exponential Smoothing and Holt-Winters models, Prediction, Forecast.

1. Introduction

The Covid-19 pandemic has created a lot of havoc in the world. It is caused by a virus called SARS-CoV-2, which comes from the family of coronaviruses and is believed to be originated from the unhygienic wet seafood market in Wuhan, China but it has now infected more than 200 countries of the world. With around 5.2 million people affected around the world (As of 22nd May, 2020, *to be updated*), it has forced people to stay in their homes and has caused huge devastation in the world economy. (Ref Singh & Jadaun [1], [2], Gupta et al. [3]).

In India, the first case of Covid-19 was reported on 30th January, which is linked to the Wuhan city of China (as the patient has travel history to the city). On 4th March, India saw a sudden hike in the number of cases and since then, the numbers are increasing day by day. As of 22nd May (*to be updated*), India has more than 124,000 cases with more than 3700 deaths. (Ref [4]).

Since the outbreak of the pandemic, scientists across the world have been indulged in the studies regarding the spread of the virus. Lin et al. [5] suggested the use of the SEIR (Susceptible-Exposed- Infectious- Removed) model for the spread in China and studied the importance of

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

government-implemented restrictions on containing the infection. As the disease grew further, Ivorra et al. [6] suggested a θ-SEIHRD model that took into account various special features of the disease. It also included asymptomatic cases into account (around 51%) in order to forecast the total cases in China (around 168500). Giordano et al. [7] also suggested an extended SIR model called SIDHARTHE model for cases in Italy which was more customized for Covid-19 in order to effectively model the course of the pandemic to help plan a better control strategy.

Petropoulos and Makridakis [8] suggested the use of exponential smoothing method to model the trend of the virus, globally. Kumar et al. [9] gave a review on the various aspects of modern technology used to fight against COVID-19 crisis.

Apart from the epidemiological models, various data-oriented models were also suggested in order to model the cases and predict future cases for various disease outbreaks from time-to-time. Various time-series models were also suggested in order to model the cases and predict future cases. ARIMA and Seasonal ARIMA models are widely used by researchers in order to model and predict the cases of various outbreaks. In 2005, Earnest et al. [10] conducted a research to model and predict the cases of SARS in Singapore and predict the hospital supplies needed using this model. Gaudart et al. [11] modelled malaria incidence in the Savannah area of Mali using ARIMA. Zhang [12] compared Seasonal ARIMA model with three other time series models to compare Typhoid fever incidence in China. Polwiang [13] also used this model to determine the time-series pattern of Dengue fever in Bangkok.

For Covid-19 as well, various researchers tried to model the cases through ARIMA. Ceylan [14] suggested the use of Auto-Regressive Integrated Moving Average (ARIMA) model to develop and predict the epidemiological trend of Covid-19 for better allocation of resources and proper containment of the virus in Italy, Spain and France. Chintalapudi et al. [15] suggested its use for predicting the number of cases and deaths post 60-days lockdown in Italy. Fanelli and Piazza [16] analyzed the dynamics of Covid-19 in China, Italy and France using iterative time-lag maps. It further used SIRD model to model and predict the cases and deaths in these countries. Zhang et al. [17] developed a segmented Poisson model to analyze the daily new cases of six countries in order to find a peak point in the cases.

Since the spread of the virus started to grow in India, various measures were taken by the Indian Government in order to contain it. A nationwide lockdown was announced on March 25th to April 14th, which was later extended to May 3rd. The whole country was divided into containment zones (where large number of cases were observed from a relatively smaller region), red zones (districts where risk of transmission was high and had higher doubling rates), green zones (districts with no confirmed case from last 21 days) and orange zones (which didn't fall into the above three zones). After the further extension of the lockdown till May 17th, various economic activities were allowed to start (with high surveillance) in areas of less transmission. Further, the lockdown is now extended to May 31st and some more economic activities have been allowed as per the transmission rates, which are the rates at which infectious cases cause new cases in the population, i.e. the rate of spread of the disease.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

On the other hand, Indian scientists and researchers are also working on addressing the issues arising from the pandemic, including production of PPE kits and tests kits as well as studying the behaviour of spread of the disease and other aspects of management. Various mathematical and statistical methods have been used for predicting the possible spread of Covid-19. The classical epidemiological models (SIR, SEIR, SIQR etc.) suggested the increasing trend of the virus and predicted the peaks of the pandemic. Early researches showed the pandemic to reach its peak by mid-May. They also showed that the basic reproduction number (R_0) and the doubling rates are lower in India, with comparison to European nations and USA. A tree-based model was proposed by Arti and Bhatnagar [18] and Bhatnagar [19] in order to study and predict the trends. They suggest that lockdown and social-distancing in India has played a significant role to control the infection rates. Chatterjee et al. [20] suggests growth of the pandemic through power law and its saturation at the later stages. Due to the complexities in the epidemic models of Covid-19, various researchers have been focusing on the data in order to forecast the future cases. Chatterjee et al. [20, 21] and Ziff & Ziff [22] suggest that after exponential growth, the total count follows a power regime of t^3, t^2, t and $\sqrt{t}$ before flattening out, where 't' refers to time. It can therefore be realized that there is an urgent need to model and forecast the growth of Covid-19 in India as the virus is in the growing stage here.

In India, the most affected states are Maharashtra with over 41,000 cases (as of 22nd May,2020 *need to be updated*), Tamil Nadu (around 14,000 cases) and Gujarat (around 13,000 cases). The greatest number of cases per million have been seen in the national capital of Delhi (621.73 cases per million). (Refer [23] for population estimates). Various states such as Arunachal Pradesh, Goa, Mizoram and Manipur have been declared Covid-19 free states as they have treated all their cases since more than 14 days. States of Nagaland and Sikkim and Union Territories of Lakshwadeep Islands and Daman and Diu are yet to report a single case. These large variations suggest the effectiveness of lockdown and sealing of state borders in containing the virus. In the latest research, Singh & Jadaun [1] studied the significance of lockdown in India and suggested that the new Covid-19 cases would stop by the end of August in India with around 350,000 total cases. While some states may see an early stopping of new cases such as Telangana (mid-June), Uttar Pradesh and West Bengal (July-end) etc., the badly affected states of Maharashtra, Tamil Nadu and Gujarat will achieve this by August-end.

Since a proven vaccine and medication is yet to be developed by the researchers then in such a scenario, modelling the present situation and forecasting the future outcome becomes crucially important in order to utilize our resources in the most optimal way. Therefore, the article aims to study the growth curve of Covid-19 cases in India and forecast its future observations. Since the disease is still in its growing age and very dynamic in nature, no model remains perfectly valid for future. We need to develop the understanding of the present situation of the pandemic.

In this article, we first study the growth curve using regression methods (exponential, linear and polynomial etc.) and propose an optimal model for fitting the cases till May 15th. Further, we propose the use of time-series models for forecasting the future observations on Covid-19 cases. Here we reach the best-fitted ARIMA model for forecasting the Covid-19 cases. We also compare these results with Exponential Smoothing (Holt-Winters) model. This study will help us

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

to understand the course of spread of SARS-CoV-2 in India better and help the government and the people to optimally use the resources available to them.

2. Statistical Methodologies

In this section, we briefly present the statistical techniques used for analyzing the Covid-19 cases in India. Here, we used usual regression (exponential, polynomial), times series (ARIMA) and exponential smoothing models.

2.1 Exponential- Polynomial Regression

Regression is a statistical technique that attempts to estimate the strength and nature of relationship between a dependent variable and a series of independent variables. Regression analyses may be linear and non-linear. A regression is called linear when it is linear in parameters e.g. $y = \beta_0 + \beta_1 t + \in$ and $y = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \in$, $\in \sim N(0, \sigma^2)$, where y is response variable, t denotes the indepenet variable, β_0 is the intercept and other βs are known as slope.

A non-linear regression is a regression when it is non-linear in its parameters e.g. $y = \beta_0^2 + \beta_1^4 t + \beta_2^2 t^2 + \in$. In the beginning of the spread of a disease, we see that the new cases are directly proportional to the existing infected cases and may be represented by $\frac{dy(t)}{dt} = ky(t)$, where k is the proportionality constant. Solving this differential equation, we get that, at the beginning of a pandemic,

$$y(t) = Ae^{kt}$$

Thus, at the beginning of a disease, the growth curve of the cases grows exponentially.

As the disease spreads in a region, governments start to take action and people start becoming conscious about the disease. Thus, after some time, the disease starts to follow a polynomial growth rather than continuing to grow exponentially.

In order to fit an exponential regression to our data, we linearize the equation by taking the natural logarithm of the equation and convert it to a linear regression in first order.

We estimate the parameters of a linear regression of order p as following-

Let the model of linear regression of order p be: $y_i = \beta_0 + \sum_{j=}^{p} x_i^j + \in$ with $\in \sim N(0, \sigma^2)$ and $i = 1,2,..,N$. Let $E = \sum_{i=1}^{N} \{y_i - \beta_0 - \sum_{j=}^{p} \beta_j x_i^j\}^2$ represent the residual sum of square (RSS).

By minimizing the RSS, we get the best estimates of these coefficients by solving the following normal equations; $\frac{\partial \mathrm{E}}{\partial \beta_0} = 0$, $\frac{\partial \mathrm{E}}{\partial \beta_1} = 0, \ldots, \frac{\partial \mathrm{E}}{\partial \beta_p} = 0$. This technique is referred to as the ordinary

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

least squares (OLS). We will use this technique of the OLS in order to estimate the coefficients of our proposed model. (Refer Montgomery et al. [24])

Since we know that the growth curve of the disease changes after some time point, exponential to polynomial, we propose to use the following joint regression model with change point μ,

$$y = \begin{cases} f_1(t); \ t \leq \ \mu, \\ f_2(t); \ t > \mu, \end{cases} \qquad \ldots(1)$$

where we take $f_1(t) = \theta_1 e^{\theta_2 t}$, $f_2(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p + \in$, $\in \sim N(0, \sigma^2)$ and p is the order of the polynomial regression model and t stands for the time (an independent variable).

During the analysis, we found that a suitable choice of $f_2(t)$ is a quadratic or a cubic model. Once the order of the polynomial is kept fixed, an optimum value of the change point can be obtained by minimizing the residuals/errors. We can obtain the OLS estimates of the parameters of the model (1) as given below.

The least square estimates (LSEs) of the parameters, $\Theta = \{\theta_1, \theta_2, \mu, \beta_0, \beta_1, \beta_2, \beta_3, \ldots \ldots, \beta_p\}$ can be obtained by minimizing the residual sum of squares (RSS) as given by-

$$\mathrm{RSS}(\Theta) = \sum_{i=1}^{\mu} \left(y_i - \hat{y}_i^{exp}\right)^2 + \sum_{i=\mu+1}^{N} \left(y_i - \hat{y}_i^{poly}\right)^2, \qquad \ldots(2)$$

where, $\hat{y}_i^{exp}$ and $\hat{y}_i^{poly}$ are the estimates value of y_i from the exponential and polynomial regression models, respectively and N is the size of the data set.

The LSEs of $\Theta = \{\theta_1, \theta_2, \mu, \beta_0, \beta_1, \beta_2, \beta_3, \ldots \ldots, \beta_p\}$ can be obtained as the simultaneous solution of the following normal equations, $\frac{\partial \mathrm{RSS}(\Theta)}{\partial \theta_1} = 0, \frac{\partial \mathrm{RSS}(\Theta)}{\partial \theta_2} = 0, \frac{\partial \mathrm{RSS}(\Theta)}{\partial \mu} = 0, \frac{\partial \mathrm{RSS}(\Theta)}{\partial \beta_0} = 0, \frac{\partial \mathrm{RSS}(\Theta)}{\partial \beta_1} = 0$, $\frac{\partial \mathrm{RSS}(\Theta)}{\partial \beta_2} = 0$, $\frac{\partial \mathrm{RSS}(\Theta)}{\partial \beta_3} = 0, \ldots \ldots, \frac{\partial \mathrm{RSS}(\Theta)}{\partial \beta_p} = 0$. Solution to these equations is difficult since the parameter μ is decenter time point. We suggest to use the following algorithm while μ is kept fixed.

Algorithm 1:

1. Set $\mu = j; \ j = 1,2, \ldots, N$.
2. For a given μ, obtain LSEs of θ_1, θ_2 using the data $\{(y_1, t_1), (y_2, t_2), \ldots, (y_j, t_j)\}$.
3. For a given μ, obtain LSEs of $\beta_0, \beta_1, \beta_2, \beta_3$ using the data $\{(y_{j+1}, t_{j+1}), (y_{j+2}, t_{j+2}), \ldots, (y_N, t_N)\}$.
4. Compute $\mathrm{RSS_j}$ using the estimates of $\{\theta_1, \theta_2, \beta_0, \beta_1, \beta_2, \beta_3, \ldots \ldots \beta_p\}$ and fixed μ
5. Repeat steps 2-4 for all $j = 1,2, \ldots, N$ and obtain the RSS for each iteration.
6. Search j^* which is a j that corresponds to the minimum RSS.
7. Take $\mu = j^*$as an optimum value of the parameter μ.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

In order to find the optimal value of μ, i.e. the turning point between the exponential and polynomial growth, we will use the technique of minimizing the residual sum squares in section 3.

We will use MAPE (Mean Absolute Percentage Error) in order to evaluate the performance of the mode.

$$\text{MAPE} = \frac{100\%}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right|,$$

where, y_t is the observed value at time point t and $\hat{y}_t$ is an estimate of y_t.

2.2 ARIMA Model

The Auto-Regressive Integrated Moving Averages method gauges the strength of one dependent variable relative to other changing variables. It is one of the most used time-series models in diverse fields of data analysis as it takes into account the changing of trends, periodic changes as well as random disturbances in the time-series data. It is used for both better understanding of the data as well as forecasting, see Brockwell et al. [25].

Autoregressive models (AR) is effectively merged with the Moving Averages models (MA) to formulate a useful time-series model, ARIMA model. The *Autoregression (AR)* element of the model shows a changing variable that regresses on its own prior values and the *Moving Average (MA)* element incorporates the dependency between an observation and a residual error from a moving average model applied to prior observations. However, this model can only be applied to stationary data. Since many real-life datasets consist of an element of non-stationarity, in order to model such datasets, ARIMA model was developed. This model is open for non-stationary data as the *Integrated (I)* factor of the model represents the differencing of raw observations to allow the time-series to become stationary.

Here, we may refer the reader to follow Box et al. [26] and Box et al. [27] for more details on ARIMA model, estimation and its application.

The general forms of AR (p) and MA (q) models can be respectively represented as the following equations:

$$Y_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \emptyset_3 Y_{t-3} + \cdots \ldots \ldots \ldots . + \emptyset_p Y_{t-p} + \varepsilon_t, \qquad \ldots(3)$$

$$Y_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \cdots \ldots \ldots \ldots . + \theta_q \varepsilon_{t-q} + \varepsilon_t, \qquad \ldots(4)$$

where ∅s and θs are autoregressive and moving averages parameters, respectively, Y_t represents value of time-series at time point t, ε_t represents the random disturbance at time point t and is assumed to be independently and identically distributed (i.i.d.) with mean 0 and variance σ^2.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The ARMA (p, q) model can be represented as-

$$Y_t = \alpha + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \emptyset_3 Y_{t-3} + \cdots \ldots \ldots \ldots . + \emptyset_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \cdots \ldots \ldots \ldots . + \theta_q \varepsilon_{t-q} + \varepsilon_t \,, \qquad \ldots(5)$$

where α is an intercept.

The differenced stationary time-series can be modelled as an ARMA model in order to use ARIMA model on the time-series data. (Refer Ceylan [14], He & Tao [28] and Manikandan et al. [29]). The ARIMA model is generally denoted as ARIMA (p, d, q) where, p is the order of auto-regression, d is the degree of difference and q is the order of moving average.

The first step to model the time-series by ARIMA is to transform the non-stationary time series into stationary time series by differencing processes. 'd' is the order of the difference. The Augmented Dickey-Fuller (ADF) Test may be applied to determine if the time series after differencing is stationary or not. The ADF test is applied to test the null hypothesis for the presence of a unit root (which indicates non-stationarity of the series).

The second step is to plot the graphs of the Autocorrelation function (ACF) and the Partial Autocorrelation Function (PACF) to determine the most-likely values of p and q.

We obtain the optimal values of p, d and q by using the AIC (Akaike Information Criterion) and CAIC (Consistent Akaike Information Criterion), for more details see https://en.wikipedia.org/wiki/Akaike_information_criterion. These information criteria may be used for selecting the best fitted models. Lower the values of criteria, higher will be its relative quality. The AIC and CAIC are given by

$$AIC = -2(\ell) + 2K,$$

$$CAIC = -2(\ell) + K\{\ln(N) + 1\},$$

where *K=number of model parameters,* $\ell = maximized\ value\ of\ log - likelihood\ function$ *and N=no. of data points.*

2.3 Exponential Smoothing

Exponential smoothing is one of the simple techniques to model time-series data where the past observations are assigned weights that are exponentially decreasing over time. We propose the following models, for modelling of Covid-19 cases (see Holt [30] and Winters [31]).

For single exponential smoothing, let the raw observations be denoted by $\{y_t\}$ and $\{s_t\}$ denote the best estimate of trend at time t. Then, $s_0 = y_0$, $s_t = \alpha y_t + (1-\alpha)(s_{t-1})$, where $\alpha \in (0,1)$ denotes the data smoothing factor.

For double exponential (Holt-Winters) smoothing, let the raw observations be denoted by $\{y_t\}$, smoothened values by $\{s_t\}$, and $\{b_t\}$ denotes the best estimate of trend at time t. Then,

$$s_1 = y_1,$$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

$$b_1 = y_2 - y_1,$$

$$s_t = \alpha x_t + (1-\alpha)(s_{t-1} - b_{t-1}),$$

$$b_t = \beta(s_t - s_{t-1)} + (1-\beta)b_{t-1},$$

where $\alpha \in (0,1)$ denotes the data smoothing factor and $\beta \in (0,1)$ denotes the trend smoothing factor. For the forecast at $t = (N+m)$ days (F_{N+m}) is calculated by

$$F_{N+m} = s_t + mb_t.$$

3. Analysis of Covid-19 cases in India

For this study, we have used the data available at GitHub, provided by Centre for Systems Science and Engineering (CSSE) at John Hopkins University (see [32]). We have used the data from March 4th to May 15th (*to be updated later*) for continuity of the data. In for this study, we use R software. (see R Core Team [33]).

3.1 Exponential-Polynomial Regressions

We know that at the beginning of the spread of the disease in India, the growth was exponential and after some time, it was shifted to polynomial. We first obtain optimum turning point of the growth, i.e. when did the growth rate of the disease shifted to polynomial regime from the exponential. We consider both quadratic and cubic regression model for second part of the data. We will also discuss the types of polynomial growth (with their equations) in India.

In order to find the turning point of the growth curve, we follow the Algorithm 1, given in the previous section. Using that, we evaluate the RSS for all the days (from March 4th) and find the date on which it is minimum. The change points of growth curve for cubic and quadratic regressions are presented in Figure 1 depending upon the size of the data set. ***From Figure 1, we can confirm that the growth rate of Covid-19 cases was exponential till April 5th and then after it follows the polynomial growth regime while we use the Covid-19 cases till May 2nd.***

Table 1: Turning point of growth curve for cubic and quadratic regression beyond change point using the Covid-19 cases from 4th March to a given day.

Day	**Change Point**	
	Cubic	Quadratic
25th April	5th April	5th April
30th April	5th April	5th April
2rd May	5th April	5th April
3rd May	7th April	5th April
5th May	10th April	11th April
10th May	10th April	18th April

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

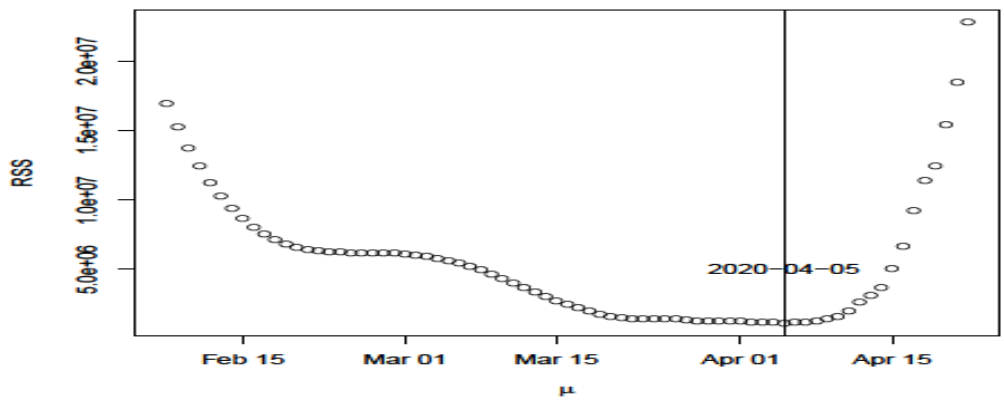

Figure 1: Trend of RSS and optimum μ for exponential-quadratic regression model.

We call the region of exponential growth in India as Region I. The coefficients of the model are presented in Table 2.

Table 2: Regression Table for Region I (Exponential Regression)

Parameter	Coefficients	S.E.	t	PV
θ_1	16.543	1.969	8.40	1.7e-09
θ_2	0.163	0.00389	41.97	2e-16

We see that after the exponential regime (till April 5th), the growth curve follows a polynomial growth till May 2nd. After this, we again see a change in the behavior of the growth curve. In Tables 3 and 4, we try to model these growth curves through regression analysis.

Table 3: Regression models fitting for Region II (5 April – 2 May).

Model	Parameter	OLS Estimates	S.E.	t	PV	R^2	(F statistic, PV)
Linear	β_0	-43427.02	1889.22	-22.99	<2e-16	0.9773	(1119, <2e-16)
	β_1	1326.49	39.66	3.45	<2e-16		
Quadratic	β_0	17410.66	1501.022	11.60	2.52e-11	0.9997	(3.901e+04, <2.26e-16)
	β_1	-1335.45	65.087	-20.52	<2e-16		
	β_2	28.32	0.6905	41.01	<2e-16		
Cubic	β_0	6196.57	10073.87	0.615	0.545	0.9997	(2.63e+04, <2e-16)
	β_1	-594.89	661.096	-0.9	0.378		
	β_2	12.29	14.2489	0.863	0.397		
	β_3	0.113	0.1009	1.126	0.272		

Table 4: Regression models fitting Table for Region III (3rd May – 15 May).

Model	Parameter	OLS Estimates	S.E.	t	PV	R^2	(F statistic, PV)
Linear	β_0	-	3244.99	-91.62	<2e-16	0.9990	(1.264e+04,

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

		2998284.88					<2e-16)
	β_1	3584.12	32.11	111.63	<2e-16		
Quadratic	β_0	-23424.15	56089.18	-0.418	0.68505	0.9997	(1.329e+04, <2.26e-16)
	β_1	-1866.15	1111.75	-1.679	0.12416		
	β_2	26.98	5.50	4.903	0.00062		
Cubic	β_0	-8.967e+05	1.837e+06	-0.488	0.637	0.9997	(1.187e+04, <2e-16)
	β_1	2.411e+04	5.464e+04	0.441	0.669		
	β_2	-2.305e+02	5.413e+02	-0.426	0.680		
	β_3	8.497e-01	1.786e+00	0.476	0.646		

Having evaluated the coefficients for various models (i.e. linear, quadratic and cubic) as well as the important statistics (i.e. R^2 values, p-values of the models as well as individual coefficients and F-statistic), we will select the best fitting models. In order to select the best fitting models for Region II (April 6th to May 2nd) and III (May 3rd to May 15th), we have the following steps. We select that model which has *high R^2 values, significant p-value, high F-statistic* and where the *p-values of all the variables are significant.*

We see for **Region II**, from Table 3 that the linear model is having a relatively lower F-statistic and R^2 values in comparison to the Quadratic and Cubic models. So, we eliminate the possibility of linear fitting. Further, we see that the p-values, F-statistics and the R^2 values are quite significant in both Quadratic as well as the Cubic models. But, if we look at the individual p-values of the coefficients, we see that the individual p-values are not significant for the Cubic model. On the other hand, the individual p-values are significant for the Quadratic model. Thus, we can conclude that the *Quadratic* model is the best fitting model for Region II (April 6th to May 2nd).

For **Region III**, from Table 4, that all the three models have high F-statistic values, high p-values and high R^2 values. But we notice that the coefficient individual p-values are not significant in both Quadratic and Cubic models. Thus, we conclude that the *Linear* model is the best fitting model for Region III (May 3rd to May 15th).

Table 5: ANOVA Table for Region II (Quadratic Regression) and III (Linear Regression)

Region	Model	Variable	Degrees of freedom	Sum of squares	Mean sum of squares	F statistic	P-Value
II	Quadratic	t	1	2882180732	2882180732	76347	$<2.2 \times 10^{-16}$
		t^2	1	63489615	63489615	1681	$<2.2 \times 10^{-16}$
		Residuals	24	906026	37751		
III	Linear	t	1	2337950722	2337950722	12462	$<2.2 \times 10^{-16}$
		Residuals	11	2063722	187611		

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Both the ANOVA tables for Region II and III suggest significant p-values for its coefficients and suggest that the models fit well the respective regions.

Thus, according to our study, the growth of the virus was exponentially increasing from March 4th to April 5th. Then after, the virus grew by following a quadratic rate from April 6th to May 2nd. Since May 3rd, we have been experiencing a linear growth, see Table 6 for best fitted regression models. Figure 2 shows the best fitted regression models to the daily cumulative cases of Covid-19 in India till 15th May.

Table 6: Course of Covid-19 growth in India

Region	Dates	Best fitted model	MAPE (%)	RSE
I	March 4th to April 5th	$y(t) = 16.54 \times e^{0.163t}$	8.60	81.66
II	April 6th to May 2nd	$y(t) = 17410.67 - 1335.45t + 28.32t^2$	0.83	194.3
III	May 3rd to May 15th	$y(t) = -29825 + 3584t$	0.49	433.1

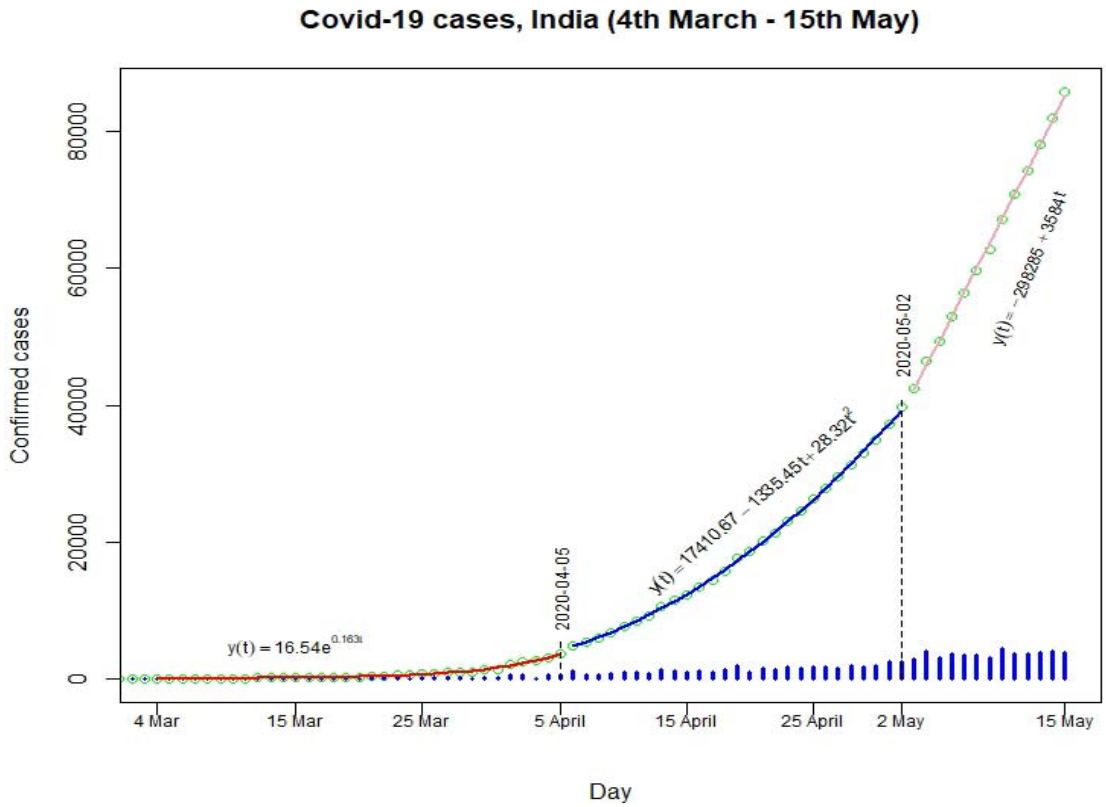

Figure 2: Fitted Regression models to the daily cumulative cases of Covid-19 in India till May 15th (Bar chart shows the daily confirmed cases).

3.2 Time series Models fitting

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

First, we check the stationarity of the transformed time-series using ADF Tests. Dickey-Fuller statistic is 11.98 with p-value 0.99 which indicates that the growth of Covid-19 cases is not stationary. The ARIMA models may be useful over the ARMA models. The ACF and PACF plots are shown in Figure 3.

We then obtain the optimal ARIMA parameters (p, d, q) by using the AIC and CAIC Criteria. We take various possible combinations of (p, d, q) and compute the AIC and CAIC Criteria. Then, select the best fitted ARIMA model that has the lowest AIC and CAIC among all considered models. According to the AIC and CAIC, the ARIMA (2, 2, 0) is the best fitted model for the Covid-19 cases, India (see Table 7). Estimates of ARIMA (2, 2, 0) parameters and MAPE are shown in Table 8.

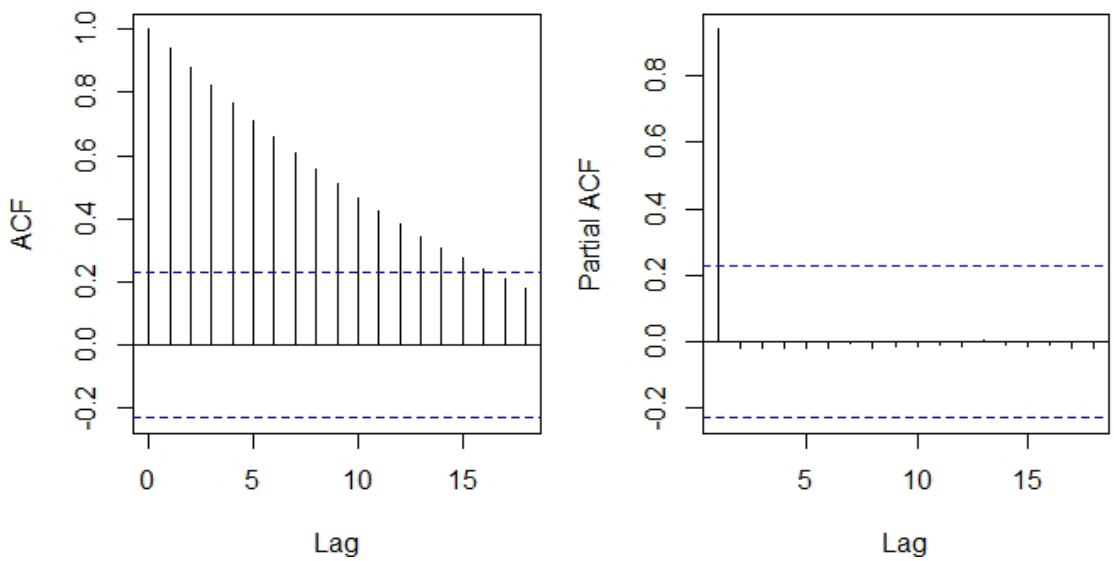

Figure 3: ACF and PACF for Covid-19 cases in India (4th March to 15th May)

Table 7: AIC and CAIC for ARIMA models for Covid-19 cases, India

Model	AIC	CAIC
ARIMA (5,2,5)	1026.64	1036.511
ARIMA (0,2,0)	1044.66	1054.531
ARIMA (1,2,0)	1025.987	1035.858
ARIMA (0,2,1)	1027.324	1037.195
ARIMA (2,2,0)	**1025.721**	**1035.592**
ARIMA (3,2,0)	1027.706	1037.577
ARIMA (2,2,1)	1027.716	1037.587
ARIMA (1,2,1)	1026.28	1036.151
ARIMA (5,1,1)	1045.42	1055.291
ARIMA (3,1,2)	1050.32	1060.191
ARIMA (3,1,1)	1049.24	1059.111
ARIMA (1,1,2)	1051.54	1061.411

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ARIMA (2,1,2)	1049.44	1059.311
ARIMA (2,1,2)	1052.41	1062.281

Table 8: Estimates of ARIMA (2,2,0) parameters and MAPE.

Coefficients	Estimate	S.E.	MAPE	Accuracy
AR 1	-0.5886	0.1160	3.87 %	96.13%
AR 2	-0.1751	0.1153		

Estimates of the Holt-Winters exponential smoothing and exponential smoothing models are given in Table 9. According to the MAPE and accuracy measures, the ARIMA (2, 2, 0) is a better model than the Holt-Winters exponential smoothing and usual exponential smoothing models. From this, we can conclude that the ARIMA model is the best fit for the cases of Covid-19, followed by Holt-Winters model. The forecasting values along with 95% confidence intervals are shown in Table 10 and Figure 4. We observe that the ARIMA model captures the trend well but it underestimates the actual Covid-19 cases. *We therefore suggest to update the ARIMA model or to use some generalized versions of the ARIMA models in future studies.*

Table 9: Estimates and MAPE of exponential smoothing models.

Model	Parameter	Estimate	MAPE	**Accuracy**
Holt-Winters Exponential Smoothing	α	0.7431282	3.93%	96.073%
	β	0.8285607		
	a	85779.225		
	b	3850.438		
Exponential Smoothing	α	0.9999527	10.33%	89.672%
	a	85783.82		

Table 10: Forecast using ARIMA and Holt-Winters models for 10 days

Day	ARIMA			Holt-Winters			Actual
	Estimate	Lower	Upper	Estimate	Lower	Upper	
16 May	89631	89000	90262	89630	89034	90226	90648
17 May	93470	92379	94561	93480	92474	94486	95698
18 May	97303	95638	98968	97331	95782	98879	100328
19 May	101141	98813	103468	101181	98994	103368	106480
20 May	104977	101923	108030	105031	102127	107936	112000
21 May	108813	104968	112658	108882	105191	112573	118224
22 May	112649	107955	117344	112732	108191	117273	
23 May	116486	110886	122085	116583	111133	122032	
24 May	120322	113767	126877	120433	114021	126845	
25 May	124158	116598	131719	124284	116857	131710	

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

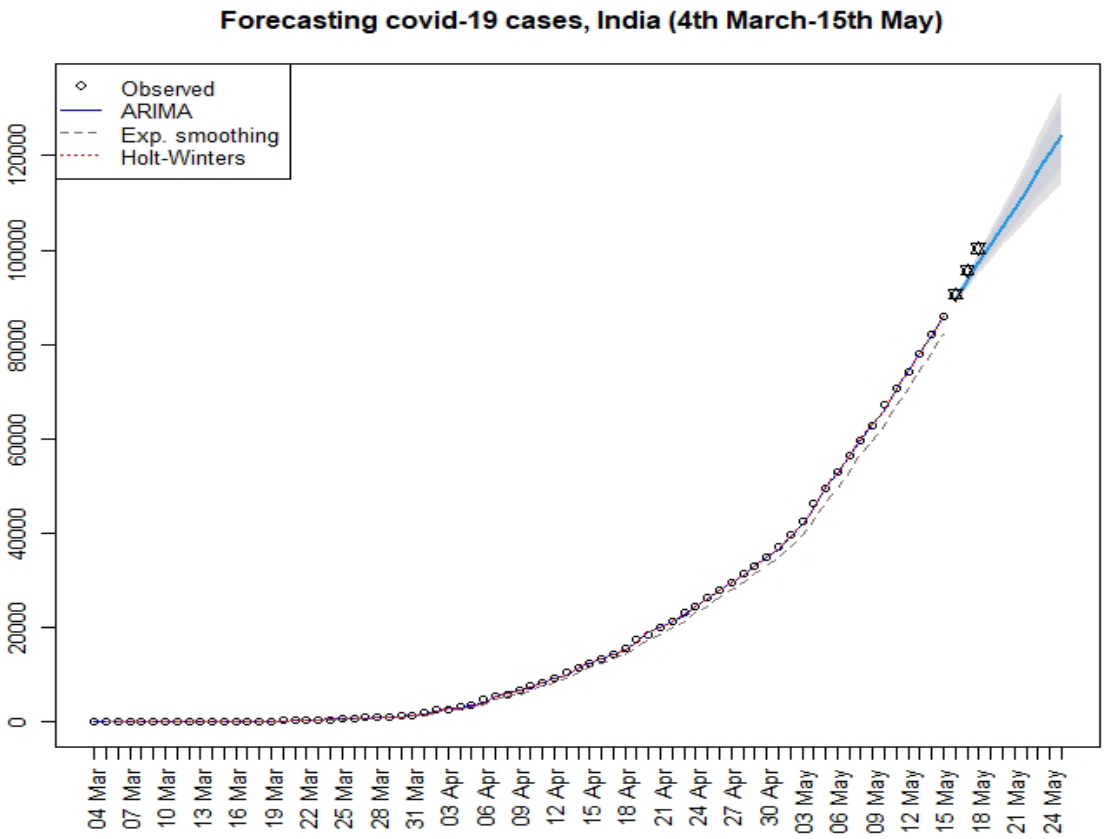

Figure 4. Fitted ARIMA (2, 2, 0) and exponential smoothing models and forecasting from ARIMA for Covid-19 cases in India (stars show the actual observations).

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

4. Conclusions

From the regression analysis, we conclude that the spread of Covid-19 disease grew exponentially from March 3rd to April 5th. Further, from April 6th to May 2nd, the cases followed a quadratic regression. From May 3rd to May 15th, we see a linear growth of the virus with average daily cases of 3584.

Verma et al. [22] showed the four stages of the epidemic, S1: exponential, S2: power law, S3: linear and S4: flat. Therefore, Covid-19 pandemic in India has entered in stage S3 of linear growth. In the days to come, it is highly likely that the total cases may start to follow a square root equation, i.e. $y(t) \sim \sqrt{t}$. And this may lead to reduction in the daily number of cases (as $y'(t) \sim 1/\sqrt{t}$,) leading to flattening of the curve.

In time series analysis, we conclude that the ARIMA (2, 2, 0) is the best fitting model for the cases of Covid-19 with an accuracy of 96.13%. The basic exponential smoothing is not very accurate for our case but we see that the Holt-Winters model is around 96.073% accurate. Both ARIMA (2, 2, 0) and Holt-Winters models suggest a rise in the number of cases in the coming days. We also observed that the ARIMA model underestimates the actual observations. Therefore, we suggest updating the ARIMA model time to time or using some generalized ARIMA models in future studies.

We may also conclude that the cases of Covid-19 will rise in the coming days and but slowly, we may head towards the reduction in the daily number of cases. But this should be accompanied by following of proper safety measures and following the guidelines of the government of India. With the gradual relaxation of lockdown measures, if proper precautions are not taken, we may see an increase in the daily cases. We must learn to lead our lives by following all the precautions once the lockdown measures are relaxed.

Acknowledgments. Dr. Vikas Kumar Sharma greatly acknowledges the financial support from Science and Engineering Research Board, Department of Science & Technology, Govt. of India, under the scheme Early Career Research Award (file no.: ECR/2017/002416).

References

1. Singh, Brijesh & Jadaun, Gunjan Singh. Modeling Tempo of COVID-19 Pandemic in India and Significance of Lockdown. medRxiv preprint doi: https://doi.org/10.1101/2020.05.15.20103325, 2020.
2. Ministry of Health and Family Welfare, Government of India, 2020. Available at https://www.mohfw.gov.in/
3. Gupta PK, Bhaskar P, Maheshwari S. Coronavirus 2019 (COVID-19) Outbreak in India: A Perspective so far. J Clin Exp Invest. 11(4):em00744. 2020.
4. https://www.worldometers.info/coronavirus/
5. Qianying Lin, Shi Zhao, Daozhou Gao, Yijun Lou, Shu Yang, Salihu S. Musa, Maggie H. Wang, Yongli Cai, Weiming Wang, Lin Yang, Daihai He, A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. International Journal of Infectious Diseases, 93, 211-216, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

6. B. Ivorra, M.R. Ferrández, M. Vela-Pérez and A.M Ramos. Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) considering its particular characteristics. The case of China. Communications in Nonlinear Science and Numerical Simulation, 88, 105303, 2020.
7. Giordano, Giulia & Blanchini, Franco & Bruno, Raffaele & Colaneri, Patrizio & Filippo, Alessandro & Matteo, Angela & Force, A SIDARTHE Model of COVID-19 Epidemic in Italy, arXiv:2003.09861 [q-bio.PE], 2020.
8. Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19. PLoS ONE 15(3): e0231236. https://doi.org/10.1371/journal.pone.0231236
9. Kumar A, Gupta PK, Srivastava A. A review of modern technologies for tackling COVID-19 pandemic, *Diabetes Metab Syndr*. 14(4):569□573, 2020.
10. Earnest, A., Chen, M.I., Ng, D., Leo, Y.S., 2005. 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, 1–8.https://doi.org/10.1186/1472-6963-5-36
11. Gaudart, J., Touré, O., Dessay, N., Dicko, A.L., Ranque, S., Forest, L., Demongeot, J., Doumbo,O.K., 2009. Modelling malaria incidence with environmental dependency in a localityof Sudanese savannah area, Mali. Malar. J. 8. https://doi.org/10.1186/1475-2875-8-61
12. Zhang, X., Liu, Y., Yang, M., Zhang, T., Young, A.A., Li, X., 2013. Comparative study of fourtime series methods in forecasting typhoid fever incidence in China. PLoS One 8.https://doi.org/10.1371/journal.pone.0063116
13. Polwiang, S. The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017). *BMC Infect Dis* **20,** 208 (2020). https://doi.org/10.1186/s12879-020-4902-6
14. Zeynep Ceylan, Estimation of COVID-19 prevalence in Italy, Spain, and France, The Total Environment Science of The Total Environment, 729, 138817, 2020.
15. Chintalapudi N, Battineni Gopi, Amenta Francesco COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach, Journal of Microbiology, Immunology and Infection, https://doi.org/10.1016/j.jmii.2020.04.004, 2020.
16. Fanelli, Duccio & Piazza, Francesco, Analysis and forecast of COVID-19 spreading in China, Italy and France, Chaos, Solitons and Fractals, Nonlinear Science, and Nonequilibrium and Complex Phenomena, 134 (2020) 109761
17. Xiaolei Zhang, Renjun Ma, Lin Wang, Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries, Chaos, Solitons and Fractals, Nonlinear Science, and Nonequilibrium and Complex Phenomena 135 (2020) 109829
18. Arti MK and Kushagra Bhatnagar, Modeling and Predictions for COVID 19 Spread in India, DOI: 10.13140/RG.2.2.11427.81444, April 01, 2020.
19. Manav R. Bhatnagar, "COVID-19: Mathematical Modeling and Predictions", submitted to ARXIV. Online available at: http://web.iitd.ac.in/~manav/COVID.pdf
20. Soumyadeep Chatterjee, B Shayak, Ali Asad, Shashwat Bhattacharya, Shadab Alam, Mahendra K. Verma, Evolution of COVID-19 pandemic: Power law growth and saturation, medRxiv 2020.05.05.20091389; doi: https://doi.org/10.1101/2020.05.05.20091389, 2020.
21. Mahendra K. Verma, Ali Asad, Soumyadeep Chatterjee, COVID-19 pandemic: Power law spread and flattening of the Curve, https://doi.org/10.1101/2020.04.02.20051680, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.20.20107540; this version posted May 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

22. A. L. Ziff and R. M. Ziff, Fractal kinetics of COVID-19 pandemic, medRxiv 2020.02.16.20023820; doi: https://doi.org/10.1101/2020.02.16.20023820, 2020.
23. https://nhm.gov.in/New_Updates_2018/Report_Population_Projection_2019.pdf
24. Montgomery, Douglas C. & Peck, Elizabeth A. & Vining, G. Geoffrey, *Introduction to Linear Regression Analysis* (Wiley, 2012)
25. Brockwell, Peter J. & Davis, Richard A., *Introduction to Time Series and Forecasting* (Springer, 1996)
26. Box, George E. P. & Jenkins Gwilym M. & Reinsel , Gregory C., *Time Analysis Analysis* (Wiley, 2008)
27. Box, George E. P. & Jenkins, Gwilym M. & Reinsel, Gregory C.& Ljung, Greta M., Time Series Analysis: Forecasting and Control (Wiley 2015)
28. He, Z., Tao, H., 2018. International Journal of Infectious Diseases Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: a nine-year retrospective study. Int. J. Infect. Dis. 74, 61–70.https://doi.org/10.1016/j.ijid.2018.07.003
29. Manikandan M, Velavan A, Singh Z, Purty AJ, Bazroy J, Kannan S. Forecasting the trend in cases of Ebola virus disease in West African countries using auto regressive integrated moving average models. Int J Community Med Public Health 2016;3:615-8
30. Holt, C. E. (1957). Forecasting seasonal and trends by exponentially weighted averages (O.N.R. Memorandum No. 52). Carnegie Institute of Technology, Pittsburgh USA. https://doi.org/10.1016/j.ijforecast.2003.09.015
31. Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*, 324–342.
32. https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
33. R Core Team. R: language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. 2020.