Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 365–383
https://www.ssca.org.in/journal

Prediction of COVID-19 Disease Progression in India under the Effect of National Lockdown

Sourish Das
Chennai Mathematical Institute, India

This is a special invited paper on request from the Chair Editor.

Received: 30 August 2024; Revised: 15 November 2024; Accepted: 20 November 2024

Abstract

In this paper, we implemented the epidemiological Susceptible-Infected-Recovered (SIR) model to estimate the basic reproduction number, $\mathcal{R}_0$, at both national and state levels in India. To the best of our knowledge, it was the first study that attempted to estimate $\mathcal{R}_0$ for India and its different states, see Das (2020). As this was the first attempt, the study used data until March 24, 2020. In the very early days of the pandemic, the data were sparse and it was difficult to conduct analysis and make forecasts. Under such circumstances, we developed a statistical machine learning model to predict future case numbers.

Our analysis showed that the situation in Punjab ($\mathcal{R}_0 \approx 16$) was critical and required immediate, aggressive intervention. We observed that the $\mathcal{R}_0$ values for Madhya Pradesh (3.37), Maharashtra (3.25), and Tamil Nadu (3.09) all exceeded 3. The $\mathcal{R}_0$ values for Andhra Pradesh (2.96), Delhi (2.82), and West Bengal (2.77) were higher than India's overall $\mathcal{R}_0$ of 2.75, as of March 4, 2020. India's $\mathcal{R}_0$ of 2.75 at that stage was comparable to that of Hubei, China during the early phase of the outbreak in December, 2019.

Our analysis indicated that India's early disease progression was similar to China. With the lockdown in place, India could have expected a number of cases comparable to, if not more than, those in China. If the lockdown had been effective, we anticipated fewer than 66,224 cases by 1 May 2020. The out-of-sample R^2 was 0.9323, and the observed number of cases on 1 May 2020 was 37,263, which was less than the predicted value, indicating the lockdown's effectiveness. All data and `R` code for this paper are available at `https://github.com/sourish-cmi/Covid19`.

Key words: Basic reproduction number; Epidemiological model; Statistical machine learning model.

AMS Subject Classifications: 62K05, 05B05

Corresponding Author: Sourish Das
Email: sourish@cmi.ac.in

1. Introduction

The World Health Organization (WHO) declared the outbreak of the novel coronavirus, COVID-19, a pandemic. It was estimated that it would take twelve to eighteen months to develop a vaccine for COVID-19 (see Ferguson *et al.* (2020)). The absence of a vaccine worsened the situation for India's already overstretched healthcare system. For example, the number of hospital beds per 1,000 population was less than one World-Bank (2021)—just one indicator of the miserable state of India's healthcare system. In the absence of a vaccine, "social distancing" was considered the optimal strategy to control the spread of the novel coronavirus Ferguson *et al.* (2020).

Aside from social distancing, widespread rapid testing and cluster testing were essential to identify infected individuals and isolate them. However, India did not have sufficient testing capacity, as widely reported in the media Biswas (2020). Although Indian scientists recently developed an affordable COVID-19 testing kit Pandey (2020), India needed a complete overhaul of its healthcare system on a war footing. In this context, India's Prime Minister Narendra Modi announced an unprecedented three-week nationwide lockdown on March 24, 2020. The purpose of the lockdown was to slow the spread of the novel coronavirus, allowing the government to pursue a multi-pronged strategy to add more beds to its hospital network, scale up production of COVID-19 testing kits, and provide personal protective equipment (PPE) for healthcare workers.

In such a grim scenario, the key question for Indian health officials was how many new confirmed cases would emerge and by what time, with the hope that the national lockdown would slow the virus's spread and buy them time to overhaul the healthcare system. However, there was uncertainty about whether the lockdown would provide the necessary slowdown of virus transmission. Even if the lockdown helped India control the virus's spread, it was not economically sustainable to extend it further, given the large number of workers employed in the informal sector as daily wage laborers. Therefore, in this policy paper, we attempted to estimate the effect of the lockdown and proposed a framework to track its effectiveness.

In this paper, we developed an epidemiological SIR model and a statistical machine learning model to predict disease progression in India. We implemented the SIR model to estimate the basic reproduction number, $\mathcal{R}_0$, at both national and state levels, to identify which states required more attention. Then, we applied the machine learning model to predict the number of cases ahead of time, so that the Indian administration could be better prepared in advance.

In Section (2), we introduced the database from which the data was downloaded and the model was built. In Section (3), we presented the methodology used to analyse and predict the data. In Section (4), we provided our analysis and prediction of the Covid-19 disease progression in India. Section (5) discusses the follow-up literature that came after this initial work, and Section (6) concluded the paper.

2. Data

In this paper, we utilised the following major databases to gather relevant data for our analysis and model development:

1. The data repository for the 2019 Novel Coronavirus, maintained by Johns Hopkins University. This globally recognized repository aggregates COVID-19 data from numerous official sources worldwide. The database is available at: `https://github.com/CSSEGISandData/COVID-19`.

2. Covid19India, a crowdsourced open-source database for India, which provides real-time updates on COVID-19 cases across Indian states and districts. This database offers a granular level of detail critical for region-specific analysis. It is available at: `https://www.covid19india.org/`.

3. Kaggle-Covid-19 in India, a dataset available on Kaggle that compiles COVID-19 data for India, including daily updates on confirmed cases, recoveries, and deaths. It also features various features like population and testing data that help enhance the predictive power of models. This dataset is available at: `https://www.kaggle.com/sudalairajkumar/covid19-in-india`.

These databases provided comprehensive and up-to-date information necessary for tracking the disease's progression and for building predictive models. By leveraging this data, we aimed to generate accurate forecasts and offer actionable insights for public health officials and policymakers.

3. Methodology

Legendary statistician Prof George Box, once said

"All models are wrong, but some are useful", see Box (1976).

Keeping this in mind, in this paper, we took a model-agnostic, two-pronged approach. The first was to understand the severity of the ground situation, and the second was to provide predictions to help health officials make informed plans. Epidemic models for infectious diseases provided insights into the dynamic behavior of disease spread. With these new insights, health officials could develop more effective intervention strategies. Moreover, such epidemic models were also used to forecast the course of the epidemic.

In addition to epidemic models, we considered statistical machine learning (SML) models, which were highly effective for prediction. Often, the interpretability of SML models was questioned. However, as we took a model-agnostic approach, we were able to use the epidemic models to understand the ground reality while adopting SML models to achieve better prediction accuracy.

3.1. SIR epidemiological model

The popular epidemic models for an infectious disease is the Susceptible, Infected, Recovered (SIR) model. The model considers a closed population. To start with, a few infected people are added to the population. It assumes that the mixing pattern is homogeneous. During the period of the sickness, the contagious people each infect on average $\mathcal{R}_0$ other people, who each then go on to infect $\mathcal{R}_0$ others, who are susceptible. The $\mathcal{R}_0$ is

popularly known as the Basic Reproduction Number. The $\mathcal{R}_0$ is the fundamental quantity of the disease progression, and higher $\mathcal{R}_0$ means, more people will tend to be infected in the course of the epidemic. The major advantage of the SIR model is it gives a number $\mathcal{R}_0$, which can be used to benchmark and compare the ground situation of different states and resource allocations can be made to those states which are hard hit. The SIR model can be described as,

$$\begin{aligned} \frac{\partial S}{\partial t} &= -\beta \frac{SI}{N} \\ \frac{\partial I}{\partial t} &= +\beta \frac{SI}{N} - \gamma I \\ \frac{\partial R}{\partial t} &= +\gamma I \end{aligned} \tag{1}$$

where S, I, and R are the number of people in the population that are susceptible, infected and recovered. The β is the transmission rate. Each susceptible person contacts β people per day; a fraction $\frac{I}{N}$ of which are infectious. Therefore $\beta \frac{SI}{N}$ move out of the susceptible group and goes into the infected group. The transmission rate is the average rate of contacts a susceptible person makes that is sufficient to transmit the infection. The parameter γ is the recovery rate, and γI is the flow out of the infected crowd and goes into the recovered group. The average duration a person spends in the infected group is $\frac{1}{\gamma}$ days. For Covid-19, $\frac{1}{\gamma}$ is around 14 days, see Ferguson *et al.* (2020).

In this paper, we followed the SIR implementation methodology as described in Towers (2012). Given $\mathcal{R}_0$, β, and γ, the implementation of the SIR model was fairly straightforward using the `deSolve` package, a solver for initial value problems of differential equations (see Soetaert *et al.* (2020)). It was known that $\mathcal{R}_0 = \frac{\beta}{\gamma}$, as noted in Brauer *et al.* (2008). We considered γ as $\frac{1}{14}$, based on Ferguson *et al.* (2020). However, we needed reliable estimates of $\mathcal{R}_0$ to implement the SIR model and predict the disease progression in India.

To estimate $\mathcal{R}_0$, we used the `R` package '`R0`', a toolbox for estimating $\mathcal{R}_0$, as described in Obadia *et al.* (2012). The time between the infection of a primary case and one of its secondary cases is referred to as the generation time, see Svensson (2007). The '`R0`' package assumed that the generation time of the infection was known and required it as input. The mean generation time for Wuhan was reported as 6.5 days Li *et al.* (2020). In this paper, we assumed the generation time followed a Gamma distribution and we estimated the mean and shape parameter of the Gamma distribution using data. Our estimated mean generation time for the Hubei province turned out to be 6.7 days, as presented in Table 2. Upon recovery from infection, we assumed that individuals were immune to re-infection in the short term, consistent with the assumption made in Ferguson *et al.* (2020).

At that time, we deployed a grid search method over the mean and shape of the Gamma distribution for the generation time process. For a particular choice of the mean (μ) and shape (κ) parameter, we generated the generation times and then, using that as input, we estimated $\mathcal{R}_0$ using the '`R0`' package in `R`. For the estimated $\mathcal{R}_0$ and γ (assumed to be 1/14), we simulated the disease progression for the period during which we observed new

incidences. We then calculated the Mean Square Error (MSE) in the following way:

$$MSE(\mu, \kappa) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{I}(t) - i_{obs}(t)\right)^2, \tag{2}$$

where $\hat{I}(t)$ was the new incidence estimated from the SIR model described in (1) at time point t, and $i_{obs}(t)$ was the actual incidence observed in the data at time point t. We estimated the mean parameter μ and shape parameter κ for which the MSE in (2) was minimized. Then, for the estimated mean and shape parameters, $\mathcal{R}_0$ was estimated using the '`R0`' package.

3.2. Statistical machine learning model

The infection rate of a typical epidemic reaches its peak and then slows down. The SIR model predicts when that peak will be reached very well because it captures the inherent dynamics of the epidemic. However, the SIR model is not as helpful for short and medium-term predictions. We also need short and medium-term predictions to forecast cases as quickly as possible so that health officials can make appropriate decisions. Statistical Machine Learning (SML) models are popular for their prediction accuracy in the short to medium term Sambasivan *et al.* (2020). Consequently, SML and SIR models complement each other.

It is important to note that SML models do not perform well in long-term prediction, particularly when predicting when the peak will be reached. With this understanding, we developed traditional SML models rather than deep learning models. We refrained from developing deep learning models because they require a large amount of data, which is not available in epidemiology. Additionally, the literature on how to apply deep learning to small datasets is still insufficient. Therefore, we focused on developing traditional regression-based SML models for short to medium-term predictions.

As different countries or provinces have varying population levels, we considered our variable of analysis to be the number of cases per 100,000 people (also known as the Rate),

$$\text{Rate} = \frac{\text{Cases}}{\text{Population Size}} \times 100,000.$$

We then modeled the Rate as a function of time, country, and time-country interaction in the following way:

$$\begin{aligned} \ln\{\text{Rate}_{it} + 1\} &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p \\ &\quad + \alpha_i + \alpha_i t + \alpha_i t^2 + \cdots + \alpha_i t^p + \epsilon, \end{aligned} \tag{3}$$

where Rate_{it} represents the Rate of the i^{th} country at the t^{th} time point, α_i is the effect of the i^{th} country, $\alpha_i t$ is the linear effect of time on the Rate of the i^{th} country, and $\alpha_i t^2$ is the quadratic effect of time on the Rate of the i^{th} country. We considered the following countries in our model: (1) India, (2) China, (3) US, (4) Iran, (5) South Korea, (6) Japan, (7) Italy, (8) France, (9) Germany, and (10) Spain.

3.3. Model training strategy for India to measure the effect of the lockdown

On March 24, 2020, India announced a national lockdown. To measure the effectiveness of the lockdown, we used all data up to March 24, 2020, to train the model and

learn its parameters. Based on the trained model, we predicted the disease progression path. Since the incubation period of COVID-19 is about 14 days, it was likely that for 14 days following the start of the lockdown, the disease would follow the predicted path and then begin to deviate downward. If the new confirmed cases fell below the predicted path, we could conclude that this was due to the effect of the lockdown. On the other hand, if the disease progression stayed on the predicted path, we would know that the lockdown did not work. If the disease progression rose above the predicted path, we could conclude that the situation worsened during the lockdown.

4. Analysis and prediction

Exploratory Data Analysis (EDA) was important for developing good predictive models. In Figure (1), we plotted the cases per 100,000 (also known as the Rate) for the US, EU, and Iran. The worst-hit regions—US, EU, and Iran—had rates in the range of 70 to 250. On the other hand, disease progression among Asian countries was very different, as shown in Figure (2). The disease progression in both India and Japan was similar. We observed an exponential rise in India and Japan, but at a much lower rate than in Western nations. China was able to flatten the curve, and South Korea managed to curb the rise from exponential to linear. However, up to that point, South Korea experienced the worst rate among the four major Asian countries.

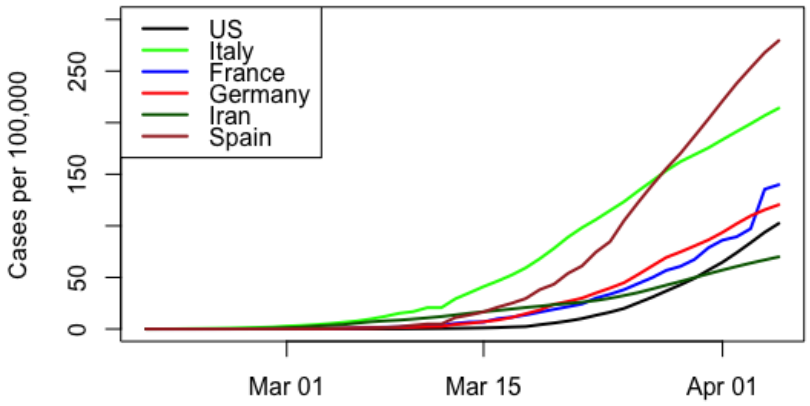


Figure 1: Cases per 100,000 in the US, EU, and Iran. The plot illustrates the rate of cases in the US, Italy, France, Germany, Iran, and Spain from early March to early April, 2020. The rates range from 70 to 250 cases per 100,000, with Spain and Italy experiencing the steepest rises, followed by Germany and the US. The data highlights the rapid escalation in Europe and Iran compared to the US during the observed period.

Prediction of Disease Progression for India from the SML model (3). The solid black vertical line in Figure (3) represented March 24, 2020. The black points to the left of

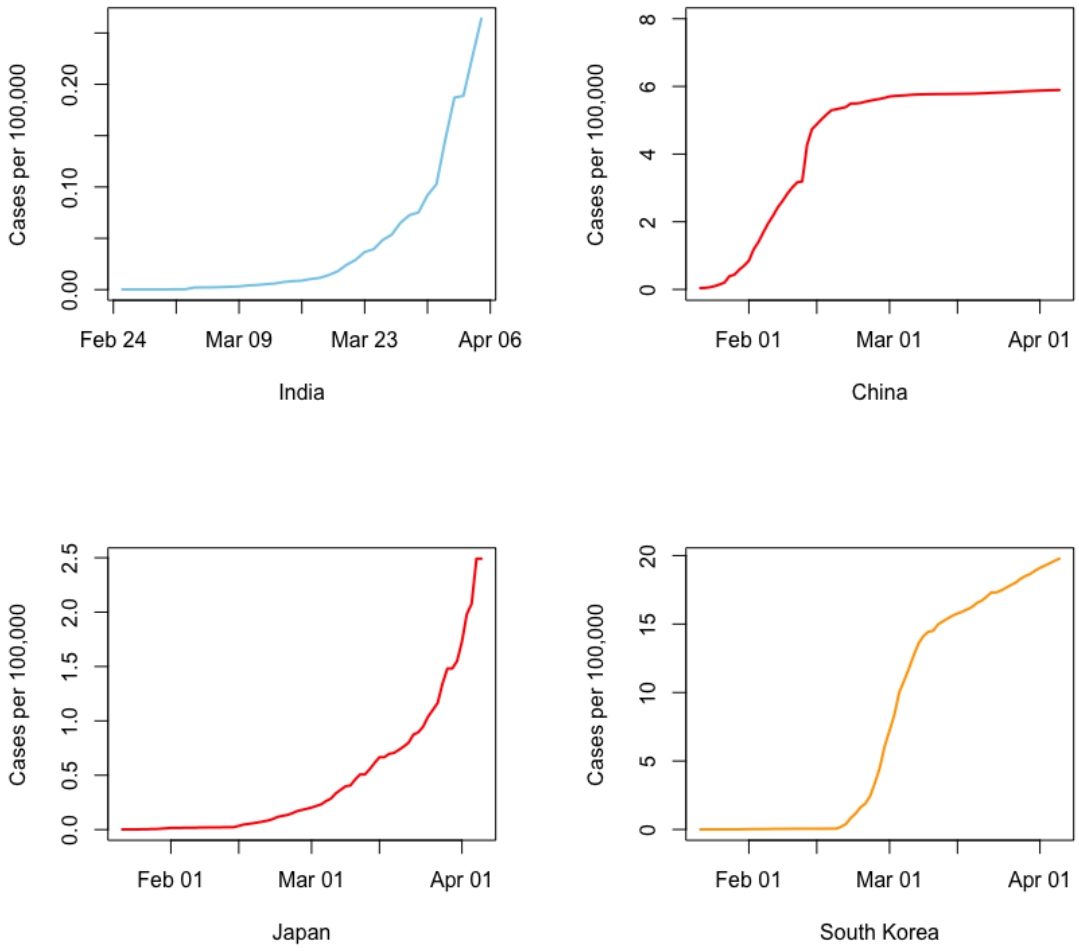


Figure 2: Cases per 100,000 in India, China, Japan, and South Korea. Note that India and Japan's cases per 100,000 are in exponential rise. However, China and South Korea were able to flatten the curve. But at different levels. China was able to flatten the curve at around 6 per 100,000 population; whereas South Korea has partially flattened its curve and increasing as a linear scale.

the vertical black line were the confirmed cases up to March 24, 2020. These black points were used in model training. The solid red line indicated the predicted path of disease progression. The blue points represented the out-of-sample test points, or the confirmed cases that appeared after March 24, 2020. As of April 7, 2020, we had not yet seen the effect of the lockdown. However, if the lockdown worked, its effect should have been visible soon. The blue points should have appeared below the predicted red line. In Table (1), we presented the actual predictions up to May 1, 2020. Had the lockdown been effective, we anticipated case numbers would stay below 66,224 by 1 May 2020. With an out-of-sample R^2 of 0.9323, the actual case count on 1 May 2020 was 37,263; below the predicted value; suggesting that the lockdown was indeed effective.

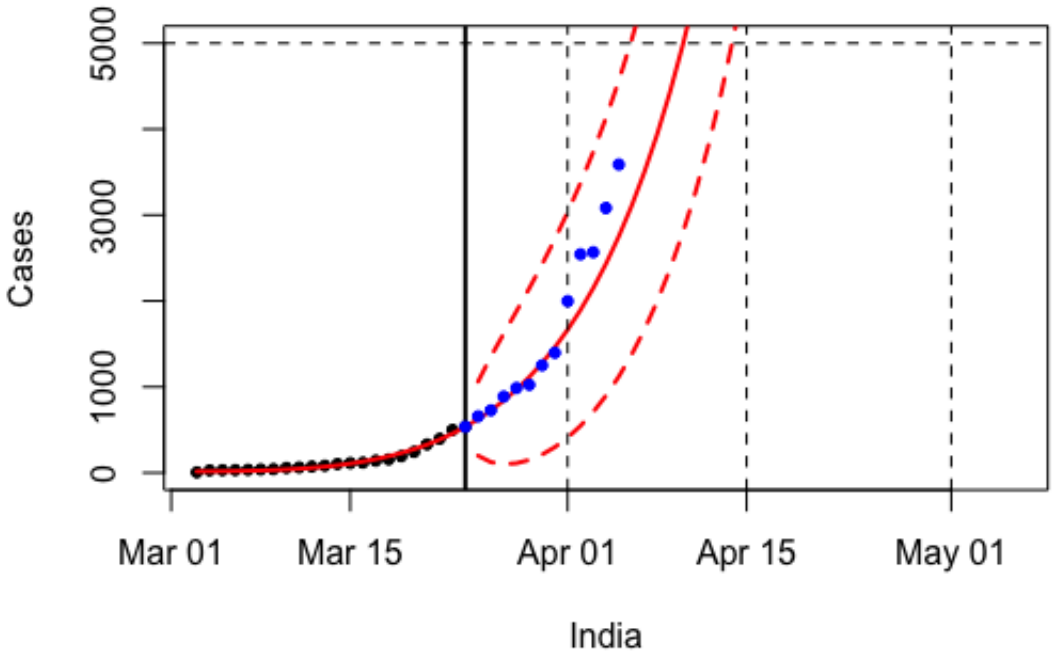


Figure 3: Predicted path of the disease progression in India. The solid black vertical line represent the 24 March 2020. The black points left of the vertical black line are confirmed cases till 24 March 2020. These black points are used in model training. The solid red line is the predicted path of the disease progression. The blue points are the out of sample test point or the confirmed cases that comes after 24 March 2020. As of 07 April 2020, we don't see the effect of lockdown. However, if lockdown works - it should shows its effect any time soon now. The blue point should appear below the predicted red line.

Comparison of $\mathcal{R}_0$ between India and China: In Table (2), the $\mathcal{R}_0$ with a 95% confidence interval for Hubei province and China was around 2.5 during the first 23 days from the start of the lockdown. India's $\mathcal{R}_0$, with a 95% confidence interval, was computed using

Table 1: The table presents the actual cases and prediction from the SML model (3). We used all the data till the 24th March 2020. Here due to space constraint, we present only 5 days interval and recent out of sample values at the daily level. Outsample $R^2 = 0.9323$. The actual values (marked in blue) were added in the current version.

	Dates	Actual Case	Prediction
1	2020-03-03	5	14.99
5	2020-03-07	34	22.42
10	2020-03-12	73	57.72
15	2020-03-17	142	158.74
20	2020-03-22	396	387.54
21	2020-03-23	499	456.29
22	2020-03-24	536	534.79
23	2020-03-25	657	624.10
24	2020-03-26	727	725.36
25	2020-03-27	887	839.85
26	2020-03-28	987	968.95
27	2020-03-29	1024	1114.20
28	2020-03-30	1251	1277.28
29	2020-03-31	1397	1460.05
30	2020-04-01	1998	1664.59
31	2020-04-02	2543	1893.20
32	2020-04-03	2567	2148.44
33	2020-04-04	3082	2433.18
34	2020-04-05	3588	2750.66
35	2020-04-06	4778	3104.50
39	2020-04-10	7599	4974.57
44	2020-04-15	12371	8838.36
49	2020-04-20	18544	15791.88
54	2020-04-25	26283	29126.81
59	2020-04-30	34867	57229.81
60	2020-05-01	37263	66223.94

two different starting points as breakouts. The first was from March 2, 2020, because the number of cases in India started rising from that day. The $\mathcal{R}_0$ for India for the first 22 days up to the lockdown was around 2.5, similar to China. However, if we used the data up to April 4, 2020, the $\mathcal{R}_0$ was around 2.75. This indicated that the situation had worsened since the lockdown, as was clear from Figure (3).

In the second approach, we considered India's breakout from January 23, 2020. In that case, if we considered the data up to March 24, 2020, the $\mathcal{R}_0$ with 95% confidence was almost 1.9, and if we considered the data up to April 4, 2020, the $\mathcal{R}_0$ was nearly 2.1. This meant that if we used the data prior to March 2, 2020, India's $\mathcal{R}_0$ appeared better. In Figure (4), we compared the incidences of Hubei and India in Figures (4:a) and (4:b). We considered the date range for Hubei from January 23, 2020, to February 14, 2020, *i.e.*, during the first 23 days of the Hubei lockdown. On the other hand, we considered the data for India from January 2, 2020, to January 24, 2020, up to the lockdown. On January 23, 2020, Hubei had

Table 2: $\mathcal{R}_0$ with a 95% confidence interval for Hubei province and China is around 2.5 during the first 23 days from the start of the lockdown. India's $\mathcal{R}_0$ with a 95% confidence interval is computed using two different starting points: one from 02-Mar-2020, as the number of cases in India started rising from that day. The $\mathcal{R}_0$ for India for the first 22 days until the lockdown is around 2.5, similar to China. However, if we use the data until 04-Apr-2020, then the $\mathcal{R}_0$ value is around 2.75. In the second approach, we consider India's breakout from 23-Jan-2020. In that case, if we consider the data until 24-Mar-2020, the $\mathcal{R}_0$ with a 95% confidence interval is almost 1.9, and if we consider data until 04-Apr-2020, the $\mathcal{R}_0$ is nearly 2.1.

	Date Range	$\mathcal{R}_0$	$\mathcal{R}_0$ Lower	$\mathcal{R}_0$ Upper	Initial Infections Considered	Mean ($\hat{\mu}$)	Shape (κ)
Hubei	23-Jan-20 to 14-Feb-20	2.53	2.50	2.57	444	6.7	0.24
China	23-Jan-20 to 14-Feb-20	2.46	2.43	2.49	548	8.7	2.7
India	02-Mar-20 to 24-Mar-20	2.52	2.35	2.71	3	5.84	6.56
India	02-Mar-20 to 04-Apr-20	2.75	2.63	2.89	3	5.41	1.10
India	23-Jan-20 to 24-Mar-20	1.87	1.78	1.97	1	2.96	1.53
India	23-Jan-20 to 04-Apr-20	2.09	2.04	2.14	1	1.25	4.98

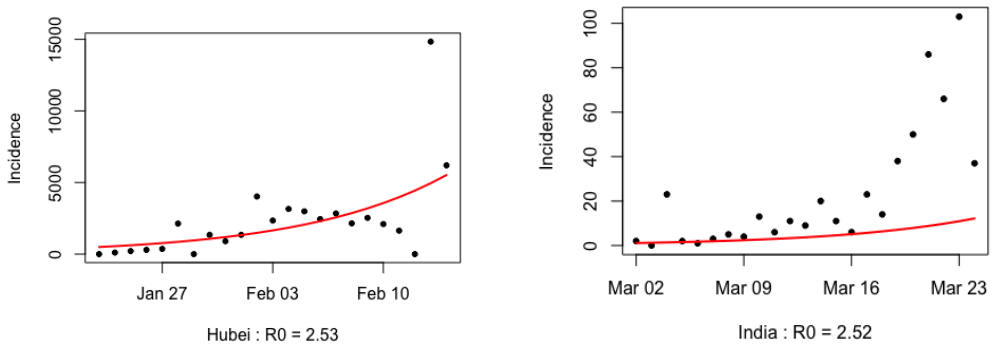
444 confirmed cases, and overall, China had 548 confirmed cases. On January 2, 2020, India had only 3 confirmed cases, whereas on the day of lockdown, *i.e.*, March 24, 2020, India had 536 confirmed cases. So, on the day when the lockdown started, both India and Hubei and/or China had a comparable number of cases.

Perhaps, we should have considered India's $\mathcal{R}_0$ to be around 2.5, similar to the early stage of COVID-19 disease progression in China. Even with the lockdown, China experienced more than 80,000 cases. Perhaps, we should have prepared for at least that many cases, if not more, in India.

State-wise $\mathcal{R}_0$: In Table (3), we presented the state-wise Basic Reproduction Number, $\mathcal{R}_0$, as of March 4, 2020. We observed that Punjab's $\mathcal{R}_0$ was the worst in the country. Punjab's high $\mathcal{R}_0 \approx 16$ was likely due to a super spreader who ignored advice to self-quarantine after returning from a trip to Italy and Germany (see BBC News (2020)). The situation in Punjab was really complicated, and serious intervention was required. In Figure (5), we presented the cases in Punjab over time. Since March 20, 2020, the number of confirmed cases increased at an unprecedented rate.

From Table (3), we saw that the $\mathcal{R}_0$ for Madhya Pradesh (3.37), Maharashtra (3.25), and Tamil Nadu (3.09) were all above 3. Clearly, the situations were complicated in these three states. The $\mathcal{R}_0$ for Andhra Pradesh (2.96), Delhi (2.82), and West Bengal (2.77) were also higher than India's overall $\mathcal{R}_0$ of 2.75. These seven states needed special attention as their $\mathcal{R}_0$ exceeded that of India. These numbers were as of April 4, 2020.

For the following states, we either did not have enough data to make inferences for $\mathcal{R}_0$, or the algorithm failed to converge: (1) Andaman and Nicobar Islands; (2) Arunachal Pradesh; (3) Chhattisgarh; (4) Goa; (5) Haryana; (6) Jharkhand; (7) Manipur; (8) Mizoram; (9) Odisha; (10) Puducherry.



(a) (b)

Figure 4: In this figure, we compare the incidences of Hubei and India in (a) and (b). We consider the date range for Hubei from 23-Jan-2020 to 14-Feb-2020, *i.e.*, during the first 23 days of Hubei lockdown. On the other hand, we considered the data for India, from the 02-Jan-2020 to 24-Jan-2020, before the lockdown. On the 23-Jan-2020, Hubei had 444 confirmed cases and overall China had 548 confirmed cases. On 02-Jan-2020, India had only 3 confirmed cases, whereas on the day of lockdown, *i.e.*, on 24-Jan-2020, India had 536 confirmed cases.

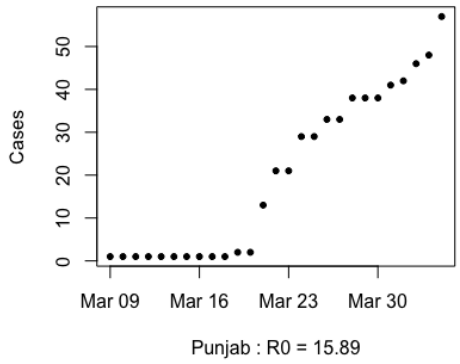


Figure 5: Confirmed cases of COVID19 in Punjab. The $\mathcal{R}_0 = 15.89$. The high $\mathcal{R}_0$ is likely due to a super spreader ignored advice to self quarantine after returning from a trip to Italy and Germany, see BBC News (2020)

Table 3: State Wise Basic Reproduction Number, $\mathcal{R}_0$, as of 04 March, 2020. Punjab's high $\mathcal{R}_0$ is likely due to a super spreader ignored advice to self quarantine after returning from a trip to Italy and Germany, see BBC News (2020)

State/UT	$\mathcal{R}_0$	Lower	Upper
Andhra Pradesh	2.96	2.56	3.45
Bihar	2.13	1.35	3.40
Chandigarh	1.14	0.89	1.48
Delhi	2.82	2.60	3.08
Gujarat	0.98	0.84	1.15
Himachal Pradesh	1.59	1.00	3.13
Jammu and Kashmir	2.02	1.69	2.48
Karnataka	2.29	1.87	2.77
Kerala	1.62	1.52	1.74
Ladakh	1.54	1.17	2.18
Madhya Pradesh	3.37	2.73	4.14
Maharashtra	3.25	2.95	3.58
Punjab	**15.89**	**4.12**	**149.27**
Rajasthan	2.45	2.25	2.67
Tamil Nadu	3.09	2.74	3.53
Telengana	2.16	1.97	2.38
Uttar Pradesh	2.30	2.10	2.52
Uttarakhand	1.33	1.13	1.61
West Bengal	2.77	2.21	3.47
India	2.75	2.63	2.89

5. Discussion

The COVID-19 pandemic has prompted extensive research to understand transmission dynamics, evaluate the impact of interventions, and forecast its trajectory. Our early-stage analysis provided a critical assessment of the severity of the situation across various Indian states. We observed that the reproduction number ($\mathcal{R}_0$) for Punjab was alarmingly high, requiring immediate and aggressive intervention. Madhya Pradesh (3.37), Maharashtra (3.25), and Tamil Nadu (3.09) also exhibited reproduction numbers above 3, indicating the need for urgent action in these states. We noted that the $\mathcal{R}_0$ values for Andhra Pradesh (2.96), Delhi (2.82), and West Bengal (2.77) exceeded India's overall $\mathcal{R}_0$ of 2.75. As of 4 March 2020, India's $\mathcal{R}_0$ was comparable to Hubei, China, during the early outbreak phase, suggesting that India could experience a similar case trajectory if effective containment measures were not implemented. Based on the assumption of lockdown efficacy, we predicted that the total cases in India might remain below 66,224 by 1 May 2020.

Subsequent studies built upon this initial analysis. Early estimates of the basic reproduction number (R_0) for India by Das (2020) placed it around 2.75, similar to China's early pandemic stage. Later, Sinha (2020) revised this estimate to approximately 1.82 by analysing time-series data of active cases in India and other countries, confirming that non-pharmaceutical interventions, such as lockdowns, were effective in reducing transmission rates but insufficient to completely halt transmission. Both Das (2020) and Sinha (2020) highlighted regional variations in COVID-19 dynamics across India. Early studies like Mittal (2020) employed Exploratory Data Analysis (EDA) to examine COVID-19 case trends in

India up to 22 April 2020, offering insights into daily and weekly case patterns, comparing trends with neighbouring and severely affected countries, and assessing India's healthcare preparedness for the pandemic.

Further descriptive studies, such as Bhatnagar *et al.* (2021), analysed COVID-19 cases in India, examining factors like age, gender, travel history, transmission type, and patient status. They found no significant correlation between age and susceptibility but observed a strong relationship between gender and transmission type. Halder *et al.* (2022) analysed mortality and recovery rates during the lockdown phases in India, revealing high correlations between active cases and both death ($R^2 = 0.8754$) and recovery rates ($R^2 = 0.9246$), though the results offered predictable insights with limited novelty.

Deo *et al.* (2020) extended the containment strategy analysis by developing a time-series SIR model to predict COVID-19 dynamics in India. Their model incorporated progressive containment measures and provided forecasts for transmission rates and daily cases under various scenarios, aligning with our early focus on timely intervention. The study by Rath *et al.* (2020) applied Linear and Multiple Linear Regression models to predict daily active COVID-19 cases in Odisha and India, achieving high accuracy (R^2 close to 1). At the state level, Tinani *et al.* (2020) explored COVID-19 modelling for hotspot states using the ARIMA model to predict cases, recoveries, and deaths across key states like Maharashtra, Delhi, and Gujarat, which corresponded with our findings on the need for focused attention on states with higher $\mathcal{R}_0$ values.

The study by Ghosh *et al.* (2020) conducted a statewise COVID-19 analysis, predicting infection trends using ensemble models and categorising states by severity to guide resource allocation, with recommended preventive measures for states with rising daily infection rates (DIR). Roy *et al.* (2021) employed ARIMA models and GIS-based spatial analysis to forecast COVID-19 prevalence in India, identifying western and southern regions as particularly vulnerable, and demonstrated ARIMA's effectiveness in epidemiological surveillance. The study by Arora *et al.* (2020) applied deep learning models, particularly LSTM variants, to predict COVID-19 case numbers in India with high accuracy (errors below 3% for daily and 8% for weekly forecasts). They categorised states into zones based on case spread and growth rates to identify hotspots, with preventive recommendations provided. Additionally, they created a website to update these predictions for authorities and researchers.

Further studies, such as Tomar and Gupta (2020), utilised data-driven methods like LSTM and curve fitting to forecast COVID-19 trends in India over a 30-day period, evaluating the effect of preventive measures and offering accurate predictions to aid health officials and administrators. Tiwari (2020) employed an SIQR model to analyse COVID-19's progression, estimating effective reproduction rates, doubling times, and infection-to-quarantine ratios, while emphasising the link between testing rates and case detection, and suggesting model enhancements for accuracy.

Recognising lockdowns' role in controlling COVID-19, Das *et al.* (2020b) proposed a Susceptible-Exposed-Infected-Recovered (SEIR) model to estimate Temporary Eradication of Spread Time (TEST) and Critical Community Size (CCS) for Indian states, supporting our initial analysis on the need for decisive action. Similarly, Kumar (2020) applied cluster analysis to identify groups within COVID-19 data across Indian states and union territories, enhancing monitoring strategies to support government and healthcare decision-making for

improved resource allocation.

Beyond epidemiology, researchers examined socioeconomic and demographic factors influencing COVID-19 outcomes. Chakravarty *et al.* (2021) analysed the impact of comorbidities, health expenditure, and life expectancy on case fatality rates across SAARC nations, underscoring the importance of targeted interventions based on local vulnerabilities, complementing our early epidemiological analysis. Broader impacts of the pandemic were explored in studies like Pyne *et al.* (2020), who assessed social vulnerabilities to guide post-pandemic recovery, particularly in India. Similarly, Dutta *et al.* (2020) analysed the economic effects of lifting or partially implementing lockdowns in Maharashtra and Gujarat, using statistical models to project future scenarios and providing additional perspectives on the socioeconomic ramifications observed in the pandemic's early stages.

Economically, Das *et al.* (2020a) examined the pandemic's effect on payment transactions in India, noting significant reductions in economic activity due to lockdowns, followed by gradual recovery in digital payments, aligning with our initial analysis of the broader economic impacts. Grover and Magan (2020) estimated Quality Adjusted Life Years (QALY) for COVID-19 patients across Indian states, offering quantitative assessments of the pandemic's impact and providing further context to our initial predictions regarding the severity of the pandemic in different regions. The study by Shruthi and Ramani (2021) analysed COVID-19's effects on financial systems, revealing that post-crisis oil market volatility impacted agricultural commodities (excluding sugar), while pre-crisis risk transmission was minimal.

Methodologically, Venkatesan (2020) addressed modelling uncertainties using back-calculation to reconstruct past infection patterns and predict future cases in India. Sarkar (2020) proposed group testing methodologies to reduce mass testing costs, particularly valuable when disease prevalence was low. These methodological refinements complemented the epidemiological insights from our early work, enhancing pandemic management approaches.

Internationally, Maleki *et al.* (2020) and Zhang *et al.* (2020) analysed COVID-19 dynamics in the U.S., with the former examining the association between comorbidities and death rates across U.S. cities, and the latter identifying change points in the pandemic's progression. These studies offered comparative insights that informed COVID-19 management in India. A different study by Gupta *et al.* (2020) investigated the influence of weather, particularly temperature and absolute humidity, on COVID-19 spread in the U.S., finding significant case increases in states with absolute humidity levels between 4 and 6 g/m^3. The results aligned with global trends and identified Indian regions potentially vulnerable to weather-driven COVID-19 transmission, underscoring weather's role in transmission risk.

In summary, our early analysis laid the foundation for subsequent research on COVID-19 in India, providing essential insights into the pandemic's progression and emphasising the need for swift, aggressive intervention in states with high $\mathcal{R}_0$ values. The extensive literature that followed expanded upon these initial findings, offering deeper insights into the epidemiological, socioeconomic, and policy dimensions of the pandemic, while validating many of our initial predictions and observations.

6. Conclusion

The conclusion of this study, aimed at predicting COVID-19 progression in India using both the SIR epidemiological model and a statistical machine learning approach, provides several key insights into the trajectory of the disease under the national lockdown. Conducted during the early phase of the pandemic, this research offers a valuable reference for understanding the dynamics of COVID-19 and implementing effective intervention strategies.

Firstly, the results underscored the critical importance of timely and aggressive interventions in mitigating the spread of COVID-19. The high basic reproduction number (R_0) observed in states such as Punjab, Madhya Pradesh, Maharashtra, and Tamil Nadu indicated the urgent need for concentrated efforts in these regions. Punjab's R_0 of 15.89 - driven by a super spreader event—highlighted the need for immediate and comprehensive containment measures. Other states, including Andhra Pradesh, Delhi, and West Bengal, also had reproduction numbers exceeding India's overall R_0, which was calculated at 2.75 as of 4 March 2020. This finding highlighted the regional disparities in COVID-19 transmission, necessitating tailored interventions to effectively curb the spread of the virus.

The study also revealed that India's disease progression mirrored that of China's early pandemic phase, particularly in terms of the R_0 values. With China's experience showing a similar reproduction number, it was evident that without a successful lockdown and containment strategy, India could have faced a similar, or even greater, number of cases. The model predicted that, if the lockdown was effective, the number of confirmed cases in India by 1 May 2020 would remain under 66,224. However, the analysis demonstrated that India's R_0 began to rise following the lockdown, indicating that while the initial lockdown slowed the virus's spread, it might not have been sufficient to halt transmission entirely.

Another significant outcome of this research was the validation of a hybrid modelling approach, where the SIR model provided accurate long-term predictions of disease dynamics, while the machine learning model excelled in short- to medium-term forecasts. This dual strategy was especially useful in understanding the immediate impacts of the lockdown, enabling public health officials to allocate resources more effectively and plan for the spread of the virus. The efficacy of the lockdown could be evaluated by comparing the actual number of cases after the lockdown to the predicted numbers based on pre-lockdown data. If new cases fell below the predicted levels, it would suggest that the lockdown was working. Conversely, if case numbers exceeded predictions, the lockdown measures would need to be reconsidered.

The use of data from multiple sources, such as Johns Hopkins University's COVID-19 repository and open-source platforms like Covid19India, was crucial in ensuring the accuracy of the model's predictions. The integration of both global and local datasets enabled a more detailed understanding of the pandemic's progression in India, a country with vast regional differences in population density, healthcare infrastructure, and socioeconomic factors. These variations were reflected in the model's predictions, which highlighted states like Kerala, with relatively lower R_0 values, indicating that local intervention efforts were somewhat successful.

However, it is important to recognise that while this study provided early estimates and predictions, the dynamics of the pandemic were rapidly evolving. Continuous data col-

lection and refinement of models would be essential to ensure that public health responses could adapt to new developments. The early prediction that India could experience over 66,000 cases by May 2020, assuming successful lockdown measures, offered a critical window for the government to expand healthcare capacity and implement more targeted interventions, such as scaling up testing, improving contact tracing, and ensuring the availability of personal protective equipment (PPE) for healthcare workers.

In conclusion, this study offered a vital early framework for understanding and predicting the spread of COVID-19 in India, delivering actionable insights for policymakers. The hybrid approach, combining epidemiological models with statistical machine learning, allowed for more accurate short- and long-term predictions, helping to shape India's pandemic response. The key takeaway is the necessity of timely, aggressive, and region-specific interventions to control the spread of infectious diseases, particularly in a country as diverse and densely populated as India. Moreover, the study emphasised the limitations of lockdowns as a long-term solution and stressed the need for a robust healthcare infrastructure and continuous policy adaptation based on real-time data.

Addendum

Prediction of Disease Progression for India: In Table (1), we presented the actual predictions up to May 1, 2020. Had the lockdown been effective, we anticipated case numbers would stay below 66,224 by 1 May 2020. With an out-of-sample R^2 of 0.9323, the actual case count on 1 May 2020 was 37,263; below the predicted value; suggesting that the lockdown was indeed effective.

Table (1) Description: The table presents the actual cases and prediction from the SML model (3). We used all the data till the 24th March 2020. The blue values were added in the current version

Acknowledgements

I sincerely thank the chief editor for inviting me to contribute this paper and for providing valuable guidance and advice. I am also deeply grateful to the reviewer for their insightful comments, helpful suggestions, and for kindly recommending numerous useful references.

Conflict of interest

The author declares no financial or non-financial conflicts of interest related to the research presented in this article.

References

Arora, P., Kumar, H., and Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons & Fractals*, **139**, 110017. Epub 2020 Jun 17.

BBC News (2020). Coronavirus: India 'super spreader' quarantines 40,000 people. `https://www.bbc.com/news/world-asia-india-52061915`. Accessed: 27 March 2020.

Bhatnagar, V., Poonia, R. C., Nagar, P., Kumar, S., Singh, V., Raja, L., and Dass, P. (2021). Descriptive analysis of COVID-19 patients in the context of India. *Journal of Interdisciplinary Mathematics*, **24**, 489–504.

Biswas, S. (2020). Coronavirus: Why is India testing so little? `https://www.bbc.com/news/world-asia-india-51922204`. Accessed: 20 March 2020.

Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, **71**, 791–799.

Brauer, F., Driessche, P. v. d., and Wu, J., editors (2008). *Mathematical Epidemiology*. Springer. Lecture Notes in Mathematics, 1945.

Chakravarty, S., Grover, G., and Aggarwal, S. (2021). Association of socioeconomic and demographic factors with COVID-19 related health outcomes in SAARC nations. *Statistics and Applications*, **19**, 367–386. ISSN 2454-7395 (online).

Das, A., Das, S., Jaiswal, A., and Sonthalia, T. (2020a). Impact of COVID-19 on payment transactions. *Statistics and Applications*, **18**, 239–251.

Das, S. (2020). Prediction of COVID-19 disease progression in India under the effect of national lockdown. First Version: April 07, 2020, Available from: `https://arxiv.org/pdf/2004.03147`.

Das, S., Ghosh, P., Sen, B., Pyne, S., and Mukhopadhyay, I. (2020b). Critical community size for COVID-19: A model based approach for strategic lockdown policy. *Statistics and Applications*, **18**, 181–196.

Deo, V., Chetiya, A. R., Deka, B., and Grover, G. (2020). Forecasting transmission dynamics of COVID-19 in India under containment measures- a time-dependent state-space SIR approach. *Statistics and Applications*, **18**, 157–180.

Dutta, S., Das, K., Chatterjee, K., and Chakraborty, A. (2020). What if lockdown is removed? district level predictions for Maharashtra and Gujarat. *Statistics and Applications*, **18**, 209–221.

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguein, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., *et al.* (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Technical report, Imperial College London, WHO Collaborating Centre for Infectious Disease Modelling, MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics. Imperial College COVID-19 Response Team, Available from `https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf`.

Ghosh, P., Ghosh, R., and Chakraborty, B. (2020). COVID-19 in India: Statewise analysis and prediction. *JMIR Public Health and Surveillance*, **6**, e20341.

Grover, G. and Magan, R. (2020). Estimation of quality adjusted life year (qaly) for different states of India during COVID-19. *Statistics and Applications*, **18**, 319–331.

Gupta, S., Raghuwanshi, G. S., and Chanda, A. (2020). Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020. *Science of The Total Environment*, **728**, 138860. Epub 2020 Apr 21; Erratum in: Sci Total Environ. 2020 Dec 15;748:142577. doi: 10.1016/j.scitotenv.2020.142577.

Halder, B., Bandyopadhyay, J., and Banik, P. (2022). Statistical data analysis of risk factor associated with mortality rate by COVID-19 pandemic in India. *Modeling Earth Systems and Environment*, **8**, 511–521.

Kumar, S. (2020). Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. *Annals of Data Science*, **7**, 417–425.

Li, Q., Guan, X., and Wu, P. e. (2020). Early transmission dynamics in Wuhan, China, of novel corona virus–infected pneumonia. *The New England Journal of Medicine*, **382**.

Maleki, M., McLachlan, G. J., Gurewitsch, R., Aruru, M., and Pyne, S. (2020). A mixture of regressions model of COVID-19 death rates and population comorbidities. *Statistics and Applications*, **18**, 295–306.

Mittal, S. (2020). An exploratory data analysis of COVID-19 in India. *International Journal of Engineering Research & Technology (IJERT)*, **9**, IJERTV9IS040550. `http://www.ijert.org`.

Obadia, T., Haneef, R., and Boëlle, P.-Y. (2012). The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making*, **12**, 147.

Pandey, G. (2020). Coronavirus: The woman behind India's first testing kit. `https://www.bbc.com/news/world-asia-india-52083196`. Accessed: 28 March 2020.

Pyne, S., Ray, S., Gurewitsch, R., and Aruru, M. (2020). Transition from social vulnerability to resiliency vis-à-vis COVID-19. *Statistics and Applications*, **18**, 197–208.

Rath, S., Tripathy, A., and Tripathy, A. R. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, **14**, 1467–1474. Epub 2020 Aug 1.

Roy, S., Bhunia, G. S., and Shit, P. K. (2021). Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. *Modeling Earth Systems and Environment*, **7**, 1385–1391.

Sambasivan, R., Das, S., and Sahu, S. K. (2020). A Bayesian perspective of statistical machine learning for big data. *Computational Statistics*, **35**, 893–930. Accepted and available here :`https://link.springer.com/article/10.1007/s00180-020-00970-8`.

Sarkar, J. (2020). Reducing the number of tests for COVID-19 infection via group testing methodologies. *Statistics and Applications*, **18**, 281–294.

Shruthi, M. and Ramani, D. (2021). Statistical analysis of impact of COVID 19 on India commodity markets. *Materials Today: Proceedings*, **37**, 2306–2311. International Conference on Newer Trends and Innovation in Mechanical Engineering: Materials Science.

Sinha, S. (2020). Epidemiological dynamics of the COVID-19 pandemic in India: An interim assessment. *Statistics and Applications*, **18**, 333–350.

Soetaert, K., Petzoldt, T., and Setzer, R. W. (2020). *deSolve: Solvers for Initial Value Problems of Differential Equations*. R package version 1.28.

Svensson, A. (2007). A note on generation times in epidemic models. *Mathematical Bioscience*, **208**, 300–311.

Tinani, K., Muralidharan, K., Deshmukh, A., Patil, B., Salat, T., and Rajodia, R. (2020). Analysis and forecasting of COVID-19 cases across hotspot states of India. *Statistics and Applications*, **18**, 223–238.

Tiwari, A. (2020). Modelling and analysis of COVID-19 epidemic in India. *Journal of Safety Science and Resilience*, **1**, 135–140.

Tomar, A. and Gupta, N. (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Science of The Total Environment*, **728**, 138762. Epub 2020 Apr 20.

Towers, S. (2012). Epidemic modelling with compartmental models using R. Available from https://sherrytowers.com/2012/12/11/simple-epidemic-modelling-with-an-sir-model/.

Venkatesan, P. (2020). A comprehensive modeling framework for estimation and prediction of COVID-19 in India. *Statistics and Applications*, **18**, 269–280.

World-Bank (2021). Hospital beds (per 1,000 people); data from the World Health Organization, supplemented by country data. license: Cc by-4.0. https://data.worldbank.org/indicator/SH.MED.BEDS.ZS?end=2021&name_desc=false&start=1960&view=chart.

Zhang, S., Xu, Z., and Peng, H. (2020). Change point modeling of Covid-19 data in the United States. *Statistics and Applications*, **18**, 307–318.