Possibility of the COVID-19 third wave in India: mapping from second wave to third wave

A Gupta and R Katarya*

Big Data Analytics and Web Intelligence Laboratory, Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

Received: 10 December 2021 / Accepted: 21 June 2022 / Published online: 14 July 2022

Abstract: After a consistent drop in daily new coronavirus cases during the second wave of COVID-19 in India, there is speculation about the possibility of a future third wave of the virus. The pandemic is returning in different waves; therefore, it is necessary to determine the factors or conditions at the initial stage under which a severe third wave could occur. Therefore, first, we examine the effect of related multi-source data, including social mobility patterns, meteorological indicators, and air pollutants, on the COVID-19 cases during the initial phase of the second wave so as to predict the plausibility of the third wave. Next, based on the multi-source data, we proposed a simple short-term fixed-effect multiple regression model to predict daily confirmed cases. The study area findings suggest that the coronavirus dissemination can be well explained by social mobility. Furthermore, compared with benchmark models, the proposed model improves prediction $R^2$ by 33.6%, 10.8%, 27.4%, and 19.8% for Maharashtra, Kerala, Karnataka, and Tamil Nadu, respectively. Thus, the simplicity and interpretability of the model are a meaningful contribution to determining the possibility of upcoming waves and direct pandemic prevention and control decisions at a local level in India.

Keywords: Coronavirus; Fixed-effect multiple regression model; Google mobility; Multi-source data; Prediction model

1. Introduction

Despite the measures taken to tackle the coronavirus pandemic, multiple epicenters have emerged all around the world. COVID-19 has wreaked havoc on India, making it one of the worst-affected countries. It has encountered approximately 34 million coronavirus-infected cases and has faced more than 450 thousand deaths across the country due to coronavirus as of October 20, 2021 [1]. Health authorities have warned that India would soon confront the third wave of coronavirus infections, though opinions vary on when it will strike the nation and how devastating it would be. Various analysts have forecasted various timelines for the coming third wave, spanning from Sept 2021 to October 2021. During the COVID-19 second wave in India, there was a drop in coronavirus cases after having the highest number of cases on May 6, 2021. A detailed examination of the pattern of reported cases in India in recent weeks indicates that the COVID-19 second wave has not yet touched its base. Since July 1, 2021, India has registered around 40,000–50,000 new cases (monthly average range), accounting for 0.12–0.15% of the overall cases. On September 1, 2021, the country reported more than 47 thousand new coronavirus cases, with five states, i.e., Kerala (32,803), Maharashtra (4456), Tamil Nadu (1509), Andhra Pradesh (1186), and Karnataka (1159) accounting for more than 88% of the total cases. Kerala and Maharashtra have seen an increase in $R_o$, having a threshold of more than 1. Many alarming findings suggest that if proper measures are not adopted quickly, the disease may spread out of control in these states.

As cases of COVID-19 decline in the mid of July 2021, the Government of India’s Ministry of Home Affairs (MHA) announced updated rules for COVID management on July 14, 2021. Due to the small number of daily COVID-19 cases, some regions have opted to allow political, social, and economic activity with caution. In addition, the states agreed to open schools, workplaces, and public transportation. However, it is considered that the second wave of coronavirus, which appeared in mid-February 2021 in India, has been fueled by people putting their guard down and attending social activities, as well as

*Corresponding author, E-mail: rahuldtu@gmail.com
conflicting messages from the government which permitted political rallies and religious gatherings.

The second wave has shown the crucial importance of focusing on pandemic preparedness. Therefore, it is necessary to analyze the status of the outbreak at the early stages of different waves of the coronavirus outbreak, focused on multi-source data. However, little knowledge about the transmission characteristics of the new variants of the disease can be obtained during its early stages. Moreover, the pandemic could be complex and fluctuate over time and space; it is challenging to accurately predict the transmission dynamics parameters; especially, tracking the social mobility effects on reducing coronavirus transmission remains unclear. With the continuous data collection and clinical research, epidemiological analysis is an important part of evaluating the consequences of mitigation selection and clinical research, epidemiological analysis is an important part of evaluating the consequences of mitigation strategies to effectively manage severely ill patients with coronavirus [2]. The availability of community mobility reports (CMR) by Google provides an opportunity to explore the association between mobility and disease cases. This report shows the changes in mobility across six categories of places: “retail and recreation,” “workplaces,” “parks,” “transit stations,” “grocery and pharmacy,” and “residential” [3]. The percentage change is calculated by comparing current mobility in certain areas to average movement before the lockdowns. These mobility patterns can be further categorized into two categories. One was mobility patterns in public places such as “retail and recreation,” “workplaces,” “parks,” “transit stations,” and “grocery and pharmacy.” The other category was mobility in residential spaces, i.e., staying at home. Given the COVID-19 pandemic’s existing transmission potential and modality, the possible impact of social mobility cannot be overlooked [4]. Research shows that a strong inoculation plan and strict social distancing protocols may help reduce COVID-19 infection peaks [5]. Since the disease spreads by direct contact, it was recommended to control the virus by reducing population contact. In a study, Reluga pointed out that social distancing is a beneficial action for reducing and controlling the transmission of the disease [6]. In 2020, Dashtbali et al. studied the social distancing effect on the transmission of contagious diseases on the network using optimal control and a differential game approach [7].

Besides human mobility, some studies have explored the relation of environmental, demographic, and socio-economic factors with coronavirus transmission. For example, Duhon et al. establish the impact of economic, climatic, demographic, and health factors on the coronavirus growth rate using the multiple linear regression model [8]. In another paper, to study the confirmed and recovered coronavirus cases globally, the Maleki et al. adopted be autoregression time series model [9]. The rapid spread of coronavirus arose the need to recognize the influence of population dynamics on the disease [10]. For example, Wehenkel observed a positive association between the elderly and the coronavirus cases [11]. Also, some studies have observed a relationship between climatic variables and coronavirus spread. Among the climatic variables, most researchers found an association between humidity, temperature, and precipitation with the survival and spread of the coronavirus [12, 13]. In addition, certain climatic conditions have been identified as strong predictors of respiratory diseases like SARS. Several statistical epidemic models have been developed to discover the effect of COVID-19 at different levels [14]. Traditional compartmental epidemic models focused on pandemic propagation dynamics are the most widely used disease models. These statistical epidemic models help in understanding the dynamics of the disease, but the model rests on simplification. For example, Wu et al. replicated the epidemic from December 31, 2019, to January 28, 2020, in mainland China using a susceptible-exposed-infectious-recovered (SEIR) model [15]. In another research, a modified SEIR model was proposed showing a gradual drop in the epidemic in China by April 2020 [16]. The author incorporated the population migration data in the proposed SEIR model. The SEIR model incorporates the incubation period of the disease by considering individuals exposed to the disease but not yet infectious. Although the SEIR model is useful for exploring transmission dynamics, it is based on simplifications. These models often run with ordinary differential equations, while the systematic non-linear dynamics models use time-varying and probabilistic parameters. Also, it is essential to outline all the parameters for estimation, such as the disease transmission rate [14]. Unfortunately, in the initial stages of the second wave of the pandemic, limited knowledge of data such as basic regeneration numbers, propagation capabilities of asymptotic cases, and so on makes appropriate forecasting of disease distribution challenging.

Due to the little knowledge at the beginning phases of COVID-19 waves, an analysis method focused on multi-source data may be used to assess and model pandemic transmission. Meanwhile, to provide plausible recommendations for the transmission risk, particularly at the initial stages of various waves of the coronavirus pandemic, we propose a short-term forecasting model focusing on population mobility data. As a study area, the research focuses on the Indian states with the maximum COVID-19 growth rate during the early phases of the second wave of the pandemic. First, we analyze multi-source data for statistics and correlations between potential variables such as population movements, environmental data, and newly confirmed cases. The findings show that the COVID-19 pandemic in the study area is well explained by human mobility. Then, a fixed-effect multiple regression model is
developed for short-term forecasting of these newly reported COVID-19 cases. Finally, we tested the effectiveness of our approach on COVID-19 data of several major cities, including Maharashtra, Kerala, Karnataka, and Tamil Nadu, from March 13, 2021, to March 27, 2021. In particular, the $R^2$ value of daily new reported cases in Karnataka and Tamil Nadu is 0.915 and 0.961, respectively. As a result, the proposed model can support the government, Disease Control Department, and public health planning in determining the possibility of an upcoming wave by analyzing the status of different factors and assessing the pandemic effectively, deploying emergency services efficiently, and organizing medical personnel to combat the epidemic. In summary, the following are the contributions of this work:

1. We propose a short-term fixed-effect multiple regression model for forecasting new daily confirmed coronavirus cases.
2. This work primarily focuses on examining the correlation between associated multi-source data and COVID-19 cases and provides evidence of their usefulness in predicting the number of daily confirmed cases.
3. We determine the factors due to which a third wave could occur.
4. We provide probabilistic predictions that estimate uncertainties in future confirmed cases.
5. We prove the efficiency of the proposed model by comparing it with the existing regression models.

This article continues as follows: Sect. 2 is devoted to the overview of the efforts to predict the COVID-19 cases with a brief introduction of the new prediction model. Section 3 is dedicated to analyzing the prediction results and discusses the contribution of the proposed model. Finally, Sect. 4 outlines the conclusion.

2. Experimental details

2.1. Data collection

Here, the multi-source data comprised the coronavirus confirmed cases, human mobility data, and environmental data for the study area were collected to determine the effect of different factors on early dynamics transmission.

2.1.1. Confirmed coronavirus cases

The almost real-time data for confirmed coronavirus cases in India were sourced from [https://api.covid19india.org/][17]. This research analyzed the data from February 20, 2021, until March 27, 2021. Figure 1 presents the time plot of the confirmed cases.

2.1.2. Human mobility data

The mobility data in public spaces beginning from February 20, 2021, until March 12, 2021, were collected from CMR by Google [3]. These data depict the flow of people who use mapping apps on their Android devices. During the study period, public spaces such as parks, workplaces, retail and recreation, and transit stations had a negative percentage change, specifying a reduction in mobility. At the same time, the positive percentage variation in mobility to places of groceries and pharmacies implies an increase in mobility in these locations [3]. These data exhibit the percentage change in daily mobility from a baseline level measured before the WHO declaration of the pandemic. According to Google, the dataset baseline is considered to be the median value for the corresponding day of the week, from January 3, 2020, to February 6, 2020.

2.1.3. Environmental data

The average temperature and humidity data were obtained from the Weather Underground portal [18], which offers the researchers a significant volume of worldwide weather data. In addition, the air quality data were collected from [https://aqicn.org][19].

2.2. Experimental design

The prevalence, dissemination, and spread of several contagious diseases are closely connected to demographics, meteorological factors, and human activities [20, 21]. In this paper, an analysis of multi-source data such as weather data and human mobility in public spaces is considered to
predict the upcoming days’ trend in coronavirus cases. Our research examines the COVID-19 growth rate for Indian states and union territories between February 20, 2021, and March 12, 2021. As a result, we take the epidemic situation of Maharashtra, the state with the maximum growth rate of coronavirus, as the primary analysis object. Further, we collect multi-source data such as average temperature, relative humidity, average concentrations of inhalable particulate matter (PM10, PM2.5), and NO2 and mobility changes in public places, and daily reported cases for the study area. Then, the correlation between the multi-source data and COVID-19 cases is examined. It is expected that the most crucial factors of pandemic propagation will be yelled out, which will help with accurate modeling. Subsequently, to render short-term predictions, we combined the proven critical factors with historical confirmed cases. The need for early and urgent warnings is satisfied in this manner.

2.3. Coronavirus growth rate

The daily growth rate of a contagious disease in a region is a key indicator of how quickly it propagates and where it reaches its peak value. The growth rate of the contagious disease is usually calculated by using exponential, logistic, and hierarchical logistic equations to fit the number of cases during the early stages of the disease [22–24]. The coronavirus disease dynamics are qualitatively close to those of a logistic curve since the number of total cases increases exponentially at first but then slows and reaches a peak. As a result, we use the logistic model and the least-squares fitting approach to approximate the COVID-19 growth rate for the second wave for each state in India [25].

A logistic growth model is utilized to fit the rate of change in the total number of confirmed coronavirus cases to the daily confirmed cases. We choose the logistic growth model to model the coronavirus pandemic as compartmental models, including the susceptible, infection, and recover (SIR) models, have a complex parametrization process and are associated with a high level of uncertainty due to the limitation of control interventions, especially in the initial stages of the pandemic. On the other hand, data-driven phenomenological models are not subject to these constraints. COVID-19 disease growth has been modeled using logistic models, employing generalized equations to account for various growth curves in different countries [26]. Wu et al. calibrated several phenomenological models to show the viability of using the logistic approach to model COVID-19 prediction [24]. Indeed, the model can provide reliable estimates of the thresholds of the coronavirus-related scenario.

The evolution of cases over time can be described by the following equation in mathematical epidemiology when using the logistic model [27]:

$$\frac{dc}{dt} = rc\left(1 - \frac{c}{K}\right)$$

(1)

where $r$ denotes the growth rate of infection; $c_t$ denotes total infection cases at time $t$, and; $K$ represents the final size of the pandemic.

The solution to the Eq. (1) is:

$$c_t = \frac{K}{1 + \left(\frac{K - c_0}{c_0}\right)\exp^{-rt}}$$

(2)

where $c_0$ is the initial number of cases.

The change in total cases is given as:

$$I_t = c_{t+1} - c_t$$

(3)

where $t$ denotes a small change in time.

We fitted the $I_t$ using the rolling 7-day average of confirmed cases. We choose the change in overall confirmed cases rather than total confirmed coronavirus cases because the cases arising from the same cumulative frequency curve are associated. All the curves, the total cases, and the change in total cases offer the same detail from a mathematical standpoint, but they do not form a statistical perspective. For cumulative curves, in which all cases from prior observations are included in the present observation [8], the assumption that the errors in observations of most curve fitting algorithms are statistically independent is incorrect.

2.4. Correlation analysis

With respect to Maharashtra, we used correlation analysis to study the relationship between human mobility, air quality, relative humidity, and average temperature and COVID-19 cases. In India, the coronavirus second wave was reported in the mid of February 2021. Since the virus needs time to incubate and analyzing the specimen takes time, it is fair to conclude that it can spread to others during this period. According to the current literature, the average incubation time is between 5 and 12 days [28–30], after which some infected people show symptoms [22, 31] while others remain asymptomatic [32]. As the prevalence of asymptomatic carriers is not accurately reported, they can become the sources of virus transmission unless some steps are taken to limit their mobility. As a result, we consider the incubation time to be 8–9 days before the onset of the disease. On a particular day $t$, in the association between the confirmed cases and the multi-source data, the number of cases may be correlated with the past lags of the latter. Since certain conditions can cause the transmission to lag, we extend the incubation period to 14 days.
In terms of various incubation periods, we investigate the relationship between the daily confirmed cases in Maharashtra from February 20, 2021, to March 12, 2021, and the associated daily google mobility data and other climatic influences. In this paper, the Pearson correlation coefficient analysis was used to investigate the relationship between multi-source data and newly confirmed coronavirus cases. The following is a description of the modified Pearson correlation:

\[
\text{pearson} = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^{n}(x_i - m_x)^2} \sqrt{\sum_{i=1}^{n}(y_i - m_y)^2}}
\]

(4)

where \(x = [x_1, x_2, ..., x_n]\) and \(y = [y_1, y_2, ..., y_n]\) represent the two vectors of length \(n\), and \(m_x\) and \(m_y\) are the means of \(x\) and \(y\), respectively.

2.5. Proposed model, fixed-effect multiple regression (FE_MR)

In this section, the proposed model generates the forecast of daily new coronavirus cases based on multi-source data and the historical confirmed cases. Unlike other conventional transmission dynamic models, our proposed model is data driven. The historical confirmed cases and the factors that correlate with the daily confirmed cases are used as inputs without loss of generality.

We consider \(F\) as the set of these variables, where \(f \in F\) represents a factor in \(F\), such as temperature, human mobility, humidity, etc. Fixing the disease incubation period, the per day value of the factor \(f\) in \(T\) time interval is denoted as:

\[
[f_{T+1}, ..., f_T]
\]

(5)

and the corresponding cumulative confirmed cases are denoted by the following sequence: \([y_{T+1}, ..., y_T]\).

Based on the above historical observations, the below equation mathematically represents the proposed model:

\[
y_{t+n} = \beta_0 + \sum_{f \in F} \beta_f f_{t-i} + \alpha_i y_{t-i-1} + W_{t+n} + u_{t-i},
\]

\(i = 1, 2, ..., T\)

(6)

where \(y_{t+n}\): number of confirmed cases on \((t + n)th\) day and \(n\) is the time step length of the factor; \(\beta_0\) is the intercept; \(\beta_f\) are the coefficients for various factors; \(\alpha_i\) represents the time-invariant, fixed-effect component; \(W_{t+n}\) is the weekend dummy; \(W_{t+n}\) has a value of 1 if \(t\) is a weekend and 0 otherwise, and \(u_{t-i}\) is a time-varying error component are the model’s trainable parameters.

To estimate the model’s parameters, we used a supervised machine learning approach.

We applied the log() transformation for confirmed cases and the response variable to achieve strictly positive predicted values. Later, we back-transformed the mentioned transformation to produce forecasts.

3. Results and discussion

This section presents the result of the growth rate estimation along with the outcome of correlation analysis between multi-source data and COVID-19 cases. Also, we present a comparison of the prediction performance of our proposed model and its benchmarks.

3.1. Evaluation criteria

The correlation analysis is conducted using the Pearson correlation coefficient and the \(p\) value. On the next prediction task, we looked at variables having a high correlation and a \(p\) value less than 0.05.

The forecasting metrics are the coefficient of determination, \(R^2\), root mean squared error (RMSE), mean absolute percentage error (MAPE), and mean square error (MAE). These matrices are calculated as follows:

\[
R^2 = \left(\frac{\sum_{i=1}^{N}(a_i - \bar{a})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^{N}(a_i - \bar{a})^2} \sqrt{\sum_{i=1}^{N}(f_i - \bar{f})^2}}\right)^2
\]

(7)

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (a_i - f_i)^2}
\]

(8)

\[
\text{MAPE} = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{a_i - f_i}{a_i} \right|
\]

(9)

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |a_i - f_i|
\]

(10)

where \(a_i\) represents the observed values and the \(f_i\) are the corresponding forecasted values on the \(i\)th day, respectively. \(\bar{a}\) and \(\bar{f}\) denote the mean value of actual and forecasted values. \(N\) represents the total number of forecast days.

These metrics help evaluate the forecast accuracy of the models from different perspectives. RMSE, MAPE, and MAE are concerned with the degree of deviation from the true value, whereas \(R^2\) shows the proportion of variance in the independent variable toward the dependent variable. A higher value of \(R^2\) and a low one of RMSE, MAPE, and MAE represent the estimation of great performance.

3.2. Growth rate estimation

We computed the growth rate for coronavirus-infected cases for all the Indian states and union territories with at
least 200 daily cases for the study period. As a result, the highest growth rates were seen in Maharashtra (1.9385) for the analysis period, followed by Kerala (1.7954), Karnataka (1.7777), and Tamil Nadu (1.7775). We performed further research focusing on data from these four regions.

3.3. Correlation between multi-source data and COVID-19 confirmed cases

Human mobility during the initial phases of the second wave of coronavirus plays a critical role in contagious pandemic transmission. A visualization of the Pearson correlation coefficient between the coronavirus cases and the observed variables with a 9-day lag is clearly seen in Fig. 2. First, it is very intuitive to notice from Fig. 2(a) that there is a significant correlation between the grocery and pharmacy mobility trends and cases of coronavirus infections, with a 9-day lag showing the greatest significance (pearsonr = 0.755, p < 0.00075). Meanwhile, the other mobility categories in Fig. 2(b)–(e) show a weak positive correlation with the new cases. It implies that the social distancing restrictions implementation in public places might effectively control the COVID-19 transmission. Noticeably, according to Fig. 2(f) and (g), the average temperature shows a negative correlation with the daily confirmed cases (pearsonr = −0.112), while relative humidity is weakly positively correlated (pearsonr = 0.665, p value = 0.001). Additionally, the results of air pollutants concentration are shown in Fig. 2(h)–(j). The findings suggest that PM2.5 shows a weak positive correlation (pearsonr = 0.364, p = 0.104), whereas PM10 and NO2 are negatively associated with reported outbreak cases (pearsonr = −0.006 and −0.285, respectively).

Table 1 summarizes the correlation coefficient between the observed factors and confirmed cases for different lag days. In order to validate if any variables have a lag effect, we have extended the interval to 14 days. The statistical findings suggest that grocery and pharmacy mobility indicates the highest consistency, while other mobility categories have a negative or weak consistency. The results indicate that, during the initial phases of the second wave of coronavirus in India, among all the mobility categories, grocery and pharmacy mobility plays a critical role in the number of daily COVID-19 cases. Moreover, it can be observed that the p values for the grocery and pharmacy mobility for every case are less than 0.05. It should be noted that grocery and pharmacy mobility has the clearest association with the daily confirmed cases at a 9-day lag effect, which corresponds with the incubation time discussed in the literature. Based on the above observation, we further perform the experiments using grocery and pharmacy mobility to verify its importance in forecasting coronavirus cases.

3.4. Short-term forecasting

As per the correlation analysis discussed in the above section, the effect of grocery and pharmacy mobility on early dissemination is chosen as the key factor for the short-term prediction of coronavirus cases. Figure 3 illustrates the time plot of the grocery and pharmacy mobility data from February 20, 2021, to March 12, 2021. Even though the time series plot is noisy, yet some systematic trends can be easily noticed. The time plot shows a pattern that the percentage change in the “grocery and pharmacy” mobility is lower on the weekend than the working days. Firstly, we only use grocery and pharmacy mobility as an input to our proposed model to further illustrate its efficacy during the initial phases of the second wave. We employ two distinct benchmark approaches to compare their prediction performance with the proposed model. The two most often used regression models for time series forecasting are Least Absolute Shrinkage And Selection Operator, LASSO [33], and Ridge Regression, RR [34]. In addition, LASSO and RR are multilinear models that can simultaneously manipulate more than one attribute [35]. These benchmark approaches are used to determine whether applying the proposed model adds value. In order to forecast the number of infected cases for Maharashtra, Kerala, Karnataka, and Tamil Nadu, the period was set from March 13, 2021, to March 27, 2021. Table 2 demonstrates the overall performance of forecasting models using $R^2$, RMSE, MAPE, and MAE.

As shown in Table 2, using only grocery and pharmacy mobility data, compared to LASSO and RR, the highest $R^2$ value is the highest achieved by our model for all the states, i.e., Maharashtra (0.671), Kerala (0.471), and Karnataka (0.857) and Tamil Nadu (0.918). Table 2 further shows that while grocery and pharmacy mobility can depict the overall trend, it is unable to portray changes in reported cases effectively. As a result, employing only one factor limits the model’s forecasting efficiency. Thus, to even further increase the forecasting accuracy, we included the historical case data. As a result, the $R^2$ value of the projection has increased to 0.789, 0.513, 0.915, and 0.961 for Maharashtra, Kerala, Karnataka, and Tamil Nadu, respectively (see Table 3), compared to only considering grocery and pharmacy mobility. Thus, it can be concluded that the
Possibility of the COVID-19 third wave in India

(a) Cases vs. grocery and pharmacy mobility
(b) Cases vs. retail and recreation mobility
(c) Cases vs. parks mobility
(d) Cases vs. transit stations mobility
(e) Cases vs. workplace mobility
(f) Cases vs. Average Temperature
(g) Cases vs. Relative Humidity
(h) Cases vs. PM2.5
(i) Cases vs. PM10
(j) Cases vs. NO2
The proposed model clearly outperforms benchmark models for all forecast accuracy measures.

In addition to the overall performance, we have also forecasted the results for the 1–5-day ahead. Table 4 reveals the Maharashtra, Kerala, Karnataka, and Tamil Nadu forecast results from 1- to 5-day ahead. Our model forecasted the confirmed cases in the above states with a high degree of precision; for some states, the $R^2$ values are greater than 0.9. Particularly for Karnataka and Tamil Nadu, the $R^2$ values of daily prediction are 0.915 and 0.961, respectively, implying that the severity of the epidemic is very well linked to the amount of human mobility in these areas. This is supported by case counts from other states in India. However, the overall trend can be captured; the $R^2$ values started to decrease as the number of predicted days $n$ increased. The following are the explanations for this phenomenon:

- As the number of prediction days, $n$, increases, there is a decrease in the number of training data for learning the model.
- The disparity in everyday mobility is still visible due to the close connection between “grocery and pharmacy” mobility and case records.

As a result, the absence of real-time data supplementation would impact the prediction model’s accuracy.

Figure 4 depicts the trends in the pandemic for Maharashtra, Kerala, Karnataka, and Tamil Nadu from March 13, 2021, to March 27, 2021. It can be easily noticed from the figure that there is a close association between the daily

---

**Table 1** Correlation coefficient between the observed factors and confirmed cases

| Variables                        | Pearson $r$ | $p$ value  |
|----------------------------------|-------------|------------|
|                                  | $n=6$       | $n=7$      | $n=8$      | $n=9$      | $n=10$     | $n=11$     | $n=12$     | $n=13$     | $n=14$     |
| GPM*                            | 0.546       | 0.581      | 0.704      | 0.755      | 0.573      | 0.582      | 0.5247     | 0.6125     | 0.5719     |
|                                  | 0.0010      | 0.005      | 0.0003     | 0.0007     | 0.0055     | 0.0056     | 0.0014     | 0.0031     | 0.0057     |
| RR*                             | 0.128       | 0.093      | 0.234      | 0.379      | 0.283      | 0.283      | 0.050      | 0.192      | 0.080      |
|                                  | 0.580       | 0.688      | 0.306      | 0.090      | 0.213      | 0.213      | 0.831      | 0.404      | 0.729      |
| Parks mobility                   | 0.145       | 0.084      | $-0.126$   | 0.104      | 0.178      | 0.240      | 0.124      | 0.161      | 0.067      |
|                                  | 0.532       | 0.718      | 0.588      | 0.653      | 0.441      | 0.294      | 0.592      | 0.487      | 0.774      |
| Transit and station mobility     | $-0.084$    | $-0.092$   | $-0.321$   | $-0.299$   | $-0.250$   | $-0.036$   | $-0.199$   | $-0.115$   | $-0.108$   |
|                                  | 0.719       | 0.691      | 0.156      | 0.188      | 0.274      | 0.875      | 0.387      | 0.620      | 0.641      |
| Workplace mobility               | 0.120       | 0.143      | $-0.099$   | 0.155      | $-0.081$   | 0.096      | 0.022      | 0.077      | 0.110      |
|                                  | 0.606       | 0.536      | 0.671      | 0.502      | 0.728      | 0.679      | 0.926      | 0.741      | 0.634      |
| PM2.5                            | 0.148       | 0.305      | 0.313      | 0.364      | 0.296      | 0.147      | 0.120      | 0.161      | 0.325      |
|                                  | 0.522       | 0.179      | 0.167      | 0.104      | 0.192      | 0.524      | 0.604      | 0.485      | 0.150      |
| PM10                             | $-0.284$    | $-0.224$   | $-0.005$   | $-0.006$   | 0.021      | $-0.049$   | $-0.220$   | $-0.240$   | $-0.203$   |
|                                  | 0.212       | 0.330      | 0.984      | 0.980      | 0.929      | 0.832      | 0.337      | 0.295      | 0.378      |
| No2                              | $-0.513$    | $-0.478$   | $-0.342$   | $-0.285$   | $-0.426$   | $-0.436$   | $-0.474$   | $-0.433$   | $-0.414$   |
|                                  | 0.017       | 0.028      | 0.129      | 0.211      | 0.054      | 0.048      | 0.030      | 0.050      | 0.062      |
| Temperature                      | $-0.097$    | $-0.022$   | $-0.112$   | 0.105      | 0.061      | 0.090      | $-0.032$   | $-0.042$   | $-0.015$   |
|                                  | 0.674       | 0.924      | 0.629      | 0.650      | 0.792      | 0.697      | 0.891      | 0.858      | 0.949      |
| Humidity                         | 0.627       | 0.550      | 0.665      | 0.504      | 0.452      | 0.472      | 0.515      | 0.585      | 0.489      |
|                                  | 0.002       | 0.009      | 0.001      | 0.020      | 0.040      | 0.031      | 0.017      | 0.005      | 0.025      |

*GPM, grocery and pharmacy mobility; RRM, retail and recreation mobility

---

![Change in GPM Percentage](image-url)
predicted cases and the actual number of cases, which demonstrates our model’s ability to forecast the early stages of a pandemic. At the beginning phases of the second wave of coronavirus in Maharashtra, our proposed model acted as a guide for disease and pandemic prediction and control.

3.5. Discussion

Our understanding of the coronavirus epidemic is constrained by numerous challenges, including gaps in the environmental factors and demographic features of cases. Likewise, nations have adopted apparently comparable yet considerably different control policies, making it challenging to appreciate which approach is effective. In this paper, we used publicly accessible datasets to show that, during the initial phase of COVID-19 second wave of the coronavirus, the dissemination of coronavirus cases in India can be explained by population mobility, especially grocery and pharmacy mobility. We utilize multi-source data to validate the spatial distribution of the second wave of the coronavirus in Maharashtra. As a result, the grocery and pharmacy mobility in Maharashtra seems to have aided in establishing transmission chains of the COVID-19 cases. Obviously, with increased human mobility, there would be a rise in the cumulative confirmed cases of coronavirus. Looking at the trends of google mobility [3], there has been an increase in grocery and pharmacy mobility since August 2021, indicating a high possibility that India will face the third wave of coronavirus if no preventive measures are taken. From the healthcare disaster experience during the second wave, India should emphasize the significance of different strategies for third-wave readiness. Also, early and massive responses from epidemic-affected regions are needed, which would significantly delay the spread of coronavirus by implementing initiatives such as lockdowns, social distancing, restrictions in grocery regions, and so on. One major limitation of this study is that we only chose to include mobility patterns and meteorological indicators, and the impact of vaccination on COVID-19 was not studied due to the lack of data at the initial stages of the second wave in India. India had a very low vaccine coverage at the initial stage of the second wave, which was increased during the second wave and may impact the magnitude and timings of the future waves in the country.

Statistical models are essential for assessing infectious disease data analyses in real-time. During the initial stages of different coronavirus waves in India, the proposed model can perform a short-term and accurate forecast of the epidemic distribution based on human mobility data. Simultaneously, it demonstrates the importance of adopting strategies such as imposing lockdown to prevent the rapid and large-scale dissemination of pandemics.

Overall, the proposed model is best suited to situations where population mobility is expected at an early stage. This model will forecast the progressive condition that will emerge in the coming days. Accordingly, effective steps can be enhanced to flatten the curve. The implementation of prevention and control policies in later stages, resulting in limited population mobility, can affect the performance.

### Table 2 One day ahead prediction performance with a single factor in different states of India

| Accuracy measure | Maharashtra | Kerala | Karnataka | Tamil Nadu |
|------------------|-------------|--------|-----------|------------|
|                  | RR | LASSO | FE_MR | RR | LASSO | FE_MR | RR | LASSO | FE_MR | RR | LASSO | FE_MR |
| $R^2$            | 0.532 | 0.503 | 0.671 | 0.555 | 0.576 | 0.471 | 0.684 | 0.640 | 0.857 | 0.763 | 0.742 | 0.918 |
| RMSE             | 3387.48 | 3561.93 | 2954.36 | 767.986 | 734.589 | 448.425 | 592.217 | 632.821 | 535.047 | 271.906 | 273.039 | 194.599 |
| MAPE             | 1.108% | 1.110% | 1.107% | 1.174% | 1.197% | 1.147% | 1.212% | 1.384% | 1.207% | 1.230% | 1.215% | 1.052% |
| MAE              | 2078.87 | 2207.53 | 1680.8 | 653.666 | 613.333 | 390.933 | 488.466 | 456.933 | 421.2 | 220.928 | 224.714 | 176.071 |

### Table 3 One day ahead prediction performance with multiple factors in different states of India

| Accuracy measure | Maharashtra | Kerala | Karnataka | Tamil Nadu |
|------------------|-------------|--------|-----------|------------|
|                  | RR | LASSO | FE_MR | RR | LASSO | FE_MR | RR | LASSO | FE_MR | RR | LASSO | FE_MR |
| $R^2$            | 0.650 | 0.630 | 0.842 | 0.637 | 0.643 | 0.713 | 0.721 | 0.718 | 0.915 | 0.829 | 0.802 | 0.961 |
| RMSE             | 843.742 | 825.034 | 770.596 | 713.771 | 556.123 | 416.024 | 727.296 | 701.897 | 420.386 | 236.990 | 238.492 | 164.313 |
| MAPE             | 1.371% | 1.592% | 1.253% | 1.29% | 1.293% | 1.272% | 1.141% | 1.155% | 1.213% | 1.165% | 1.152% | 1.179% |
| MAE              | 718.467 | 710.6 | 604.333 | 553.6 | 429.866 | 326.928 | 637.214 | 607.214 | 346.357 | 203.285 | 206.857 | 139.285 |
of the short-term model to a certain degree. However, in the early stages, with stable and large-scale population mobility, the possibility of surpassing imported cases by local population transmission cannot be ignored. Therefore, to determine the risk of disease spread in a stable local scenario, it is required to examine the existing local status.

### Table 4

| Region            | Accuracy measure | n = 2   | n = 3   | n = 4   | n = 5   |
|-------------------|------------------|---------|---------|---------|---------|
| Maharashtra       | $R^2$            | 0.791   | 0.754   | 0.622   | 0.496   |
|                   | RMSE             | 9113.093| 10,047.713| 12,666.566| 16,558.677|
|                   | MAPE             | 38.54% | 40.24% | 38% | 55% |
|                   | MAE              | 6308.733| 6544.467| 7653 | 9871.667|
| Kerala            | $R^2$            | 0.577   | 0.555   | 0.541   | 0.428   |
|                   | RMSE             | 583.540| 423.650 | 474.220 | 1076.430|
|                   | MAPE             | 27.45% | 16.44% | 24.43% | 58.36% |
|                   | MAE              | 246.62 | 318.82 | 321.21 | 528.46 |
| Karnataka         | $R^2$            | 0.901   | 0.905   | 0.826   | 0.501   |
|                   | RMSE             | 478.320| 693.440 | 361.830 | 761.820 |
|                   | MAPE             | 23.76% | 39.72% | 14.73% | 32.81% |
|                   | MAE              | 322.71 | 313.71 | 145.3 | 430.74 |
| Tamil Nadu        | $R^2$            | 0.956   | 0.868   | 0.772   | 0.510   |
|                   | RMSE             | 177.870| 341.470 | 431.270 | 612.870 |
|                   | MAPE             | 17.33% | 30.13% | 25.59% | 33.62% |
|                   | MAE              | 114.93 | 237.12 | 278.15 | 473.85 |

Fig. 4 Number of cases estimated for (a) Maharashtra, (b) Kerala, (c) Karnataka and (d) Tamil Nadu
of the disease. Furthermore, by carefully assessing the multidimensional variables such as climatic data, human behavior, and so on, a regional transmission risk model can be developed to study the development of contagious disease dynamics within the community, as well as more scientific and realistic recommendations for everyday protection of citizens and decision-makers.

4. Conclusions

This paper is intended to examine the relationship between daily confirmed cases and certain potential factors using multi-source data during the initial phases of the second wave of coronavirus in India. We found that during the early stage of the coronavirus wave, grocery and pharmacy mobility has the highest correlation with the dissemination of the coronavirus pandemic. Therefore, it can be concluded that mobility at grocery and pharmacy places should be tracked to prevent the coronavirus disease from spreading. Based on mobility data and COVID-19 cases, we suggested a fixed-effect multiple regression model for short-term estimation of the coronavirus epidemic. The model will offer robust solutions for policymakers and decision-makers by tracking human mobility in real-time. As a result, they will be able to determine the possibility of upcoming COVID-19 waves and can develop adequate measures timely and save as many lives as possible. The future work will focus on including the impact of vaccination on the spread of COVID-19. Also, we intend to explore other statistical techniques such as Structured Equation Modeling for analyzing the significant relationships between observed and latent variables.

Acknowledgements This Research work is done under DST project entitled “TRIP AID - Travel Risk Information in a Pandemic using Artificial Intelligence Development” under Ministry of Science & Technology, Department of Science & Technology, National Council for Science & Technology Communication Division (NCSTC), Health and Risk Communication Program ’Year of Awareness on Science and Health (YASH) for COVID19, Government of India, Technology Bhavan New Delhi 110016, Under File No: CO/B/FP/G110/2020 of dated January 8, 2021 by Principal Investigator (Dr. Rahul Katarya).

Funding No financial support was received from any institution or organization for this study.

Declarations

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

References

[1] WHO World Heal. Organ. 1 (2021).

[2] R M Anderson, H Heesterbeek, D Klinkenberg and T D Hollingsworth Lancet 395 931 (2020).

[3] Google Community Mobility Reports, https://www.google.com/covid19/mobility/.

[4] Y Pan, A Darzi, A Kabiri, G Zhao, W Luo, C Xiong and L Zhang Sci. Rep. 10 1 (2020).

[5] K K F Li, S A Jarvis and F Minhas Comput. Biol. Med. 134 104369 (2021).

[6] T C Refuga PLoS Comput. Biol. 6 1 (2010).

[7] M Dashthali, A Malek and M Marzaie Optim. Control Appl. Methods 41 2149 (2020).

[8] J Duhon, N Bragazzi and J D Kong Sci. Total Environ. 760 144325 (2021).

[9] M Maleki, M R Mahmoudi, D Wraith and K H Pho Travel Med. Infect. Dis. 37 101742 (2020).

[10] J B Dowd et al. Proc. Natl. Acad. Sci. U. S. A. 117 9696 (2020).

[11] C Wehenkel PeerJ 8 e10112 (2020).

[12] M M Menebo Sci. Total Environ. 737 139659 (2020).

[13] L C Chien and L W Chen Stoch. Environ. Res. Risk Assess. 34 1675 (2020).

[14] S R Vadyala, S N Betgeri, E A Sherer and A Antrimphale Array 11 100085 (2021).

[15] J T Wu, K Leung and G M Leung Lancet 395 689 (2020).

[16] Z Yang et al. J. Thorac. Dis. 12 165 (2020).

[17] COVID19-India API | api COVID19-India API | api, https://api.covid19india.org/.

[18] Weather Underground Local Weather Forecast, News and Conditions | Weather Underground, https://www.wunderground.com/.

[19] Aqicn.org Air Pollution in India, https://aqicn.org/city/india/.

[20] R Labonté, K Mohindra and T Schrecker Annu. Rev. Public Health 32 263 (2011).

[21] V Charu, S Zeger, J Gog, O N Bjørnstad, S Kissler, L Simonsen, B T Grenfell and C Viboud PLoS Comput. Biol. 13 e1005382 (2017).

[22] V Moorothy, A M H Restrepo, M P Preziosi and S Swaminathan Bull. World Health Organ. 98 150 (2020).

[23] Z Liu, P Magal, O Seydi and G Webb Math. Biosci. Eng. 17 3040 (2020).

[24] K Wu, D Darcet, Q Wang and D Sormente Nonlinear Dyn. 101 1561 (2020).

[25] R Haberman Mathematical Models: Mechanical Vibrations, Population Dynamics, and Traffic Flow (Society for Industrial and Applied Mathematics) (1999).

[26] A L Bertozzi, E Franco, G Mohler, M B Short and D Sledge Proc. Natl. Acad. Sci. U. S. A. 117 16732 (2020).

[27] F Brauer Infect. Dis. Model. 2 113 (2017).

[28] W Dhoubi, J Maatoug, I Ayouni, N Zammit, R Ghammem, S Ben Fredj and H Ghannem Syst. Rev. 10 1 (2021).

[29] J A Backer, D Klinkenberg and J Wallinga Eurosurveillance 25 200062 (2020).

[30] L N Nottmeyer and F Sera Environ. Res. 196 110977 (2021).

[31] H A Rothan and S N Byrareddy J. Autoimmun. 109 102433 (2020).

[32] M Day BMJ 368 m1165 (2020).

[33] R Tibshirani J. R. Stat. Soc. Ser. B 58 267 (1996).

[34] A E Hoerl and R W Kennard Technometrics 12 55 (1970).

[35] W Zh, Zou The COVID-19 Pandemic Prediction in the US Based on Machine Learning (Institute of Electrical and Electronics Engineers Inc.) p 283 (2020).

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