2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/Confluence51648.2021.9377115

Analyzing and Forecasting COVID-19 Outbreak in India

Shreyansh Chordia
Department of Computer Engg.
Veermata Jijabai Technological Institute
Mumbai, India 400019
smchordia_b17@ce.vjti.ac.in

Yogini Pawar
Department of Computer Engg.
Veermata Jijabai Technological Institute
Mumbai, India 400019
ygpawar_b17@ce.vjti.ac.in

***Abstract*—The unprecedented outbreak of the COVID-19 virus has infected more than 50 million people all over the world in less than a year. More than 1 million people have lost their lives due to the ongoing pandemic. The pandemic struck India on January 30, 2020, when the first positive case of COVID-19 was identified in Kerala. Today, India is one of the most adversely affected countries in the world. Hence, it is of utmost importance to analyze the trends in India and use the adopted knowledge to forecast the future course of outcomes. Along with the overall trend analysis in India, this study also takes into account 5 most affected states of the country: Maharashtra, Andhra Pradesh, Tamil Nadu, Karnataka and Uttar Pradesh as the subjects of the research. ARIMA and Prophet time series forecasting models have been used to make three types of predictions: confirmed cases, deaths and recovered cases in India as well as in the adopted states. The effectiveness of the forecasting models is evaluated based on metrics such as Root Mean Squared Error, Mean Absolute Error, Mean Absolute Percentage Error and Coefficient of Determination. The results suggest that the adopted models are promising mechanisms for forecasting COVID-19 trends. Our study also suggests that ARIMA model performs better than Prophet Model at this task of forecasting the outbreak. The forecasts can be useful in increasing the preparedness level of government authorities, health facilities and hospitals to combat against massive spread of the virus.**

***Index Terms*—COVID-19, India, ARIMA, Prophet**

I. Introduction

The COVID-19 pandemic in India is a part of the worldwide outbreak of coronavirus disease, which originated in Wuhan, China as reported in December 2019. According to the World Health Organization (WHO), the virus can range from the common cold to the Middle East respiratory syndrome coronavirus and the Severe Acute respiratory syndrome coronavirus [1]. The first COVID-19 case in India was reported in the state of Kerala, on January 30, 2020. Since then, there has been an immense increase in the number of cases in the world and eventually, India became one of the worst-affected countries. Looking at these increasing cases, a nationwide lockdown was imposed by the Prime Minister of India on March 25, 2020. This situation has affected the citizens of the country in various fields, like education, agriculture, economy, trade, defense, entertainment and various other services and has left the country in a precarious state.

The spread of COVID-19 can be classified [4] [3] under three major stages-

1. *Local outbreak:* In this stage, the chain of spread can be tracked and the source of infection can be searched for. Generally in this stage, cases are mostly within families, friends, or local exposure.

2. *Community transmission:* In this stage tracking down the source of infection isn't possible anymore. From here onwards, cases start showing up in clusters and conditions like mass spread among localities start becoming common.

3. *Large scale transmission:* In this stage the spread becomes uncontrollable and starts spreading among different regions and countries. This happens because the movement and travel of people cannot be controlled on a large scale.

Although trials are going on, no vaccine candidates are ready yet. In such a situation, it is quite essential to look over the numbers and understand the gravity of the situation. The main objective of this paper is to analyze the current trends of the COVID-19 outbreak in India and build forecasting models so that the future patterns could be examined and the preventive measures could be taken accordingly. State-wise analysis has also been done in the 5 worst-affected states of the country. Thus, we get a quantitative analysis of the outbreak in the country and the future predictions would be useful to increase the preparedness level to combat the pandemic in an even more effective way.

The algorithms used for forecasting are *ARIMA Forecasting Model* and the *Prophet Model by Facebook*. We present the trend analysis using performance metrics like RMSE, MAE, MAPE and R^2. It must be noted that the scope of this paper is limited to research on the COVID-19 outbreak, only in India and not worldwide. Also, the data [18] that has been used for our research has recorded information right from the day when India confirmed its first COVID-19 case in Kerala on 30th January 2020, latest up to 7th October 2020.

II. Related work

After the massive outbreak of COVID-19 pandemic in the world, large scale research and mathematical modeling have been carried out to understand the trends and take various preventive measures accordingly.

For analytical purposes, a lot of forecasting has been done using time series forecasting models like ARIMA and Exponential Smoothing [5] [6] [7]. These techniques are used

widely for various predictions and forecasts on time-series information within a quick time. The forecasting has also been done using regression models like Linear Regression, LASSO Regression and Support Vector Machine [8] and these models have been evaluated using parameters like R^2 score, adjusted R^2 score, Mean Absolute Error and Mean Square Error.

S. S. Arun et. al [9] have proposed well-known machine learning techniques as well as mathematical modeling techniques such as Rough Set Support Vector Machine (RS-SVM), Bayesian Ridge and Polynomial Regression, SIR model and RNN to observe the transmission of the disease and predict the scale of the pandemic.

The transmission rate of the pandemic has been calculated by several scientists and researchers using the Recurrent Neural Networks [11]. This paper forecasts the number of confirmed cases by training a set of Recurrent Neural Networks(RNN) using features like transmission rate along with meteorological factors like temperature and humidity. In [12], the authors have used LSTM network for modeling time series and compared the epidemic transmission rates in Canada, with Italy and the USA. Heni Bouhamed compiled the pandemic data of 79 countries between the date of their first case and March 13, 2020 and used Long Short-Term Memory (LSTM) architecture for the incessant observation of the disease [13].

Prediction of COVID-19 in the USA has been done in [14] using an adaptive modeling/estimation strategy based on the use of concatenated *Riccati-type modules*(each described by a parabolic phase-space representation) and suitable adaptive statistical estimation methods. This method has provided valuable insights into the dynamic characteristics of the infectious process.

Susceptible-Infected-Recovered(SIR) [19] models have been used to make short and long-term predictions on daily basis. Studies have also been done using Prophet model by Facebook and predicted the number of cases in the upcoming months in India [10].

Several other kinds of research have also been done for forecasting the spread of COVID-19 in India [7] by extending the SIR model to SIRD (Susceptible-Infectious-Recovered-Death) model [15] [16]. Another research adopted the Bhilwara model of containment and used SEIR (Susceptible-Exposed-Infectious-Removed) Model for analyzing the disease spread [17]. This paper also aims to study the spread of COVID-19 patterns in India using *ARIMA* and *Prophet Models* for forecasting.

III. Data Visualisation

The dataset [18] that has been adopted for understanding and analyzing the spread of COVID-19 in India records day-to-day state-wise COVID-19 statistics in India right from 30th January 2020 when the first case of infection was identified in Kerala state, latest till 7th October 2020.

Before forecasting, the nature of the spread, its trend and pattern needs to be understood. Hence before applying forecasting algorithms on the available data we visualize the spread throughout the country using several plots that have been built using Plotly Library in Python.

We have studied and analyzed new cases, confirmed cases, active cases, recovered cases, death cases, mortality rate, recovery rate and daily infection rate for India as well as for the adopted states. These states account for more than 50% of the total confirmed cases in India. Therefore, the study of these states along with the country as a whole provides a better understanding of the spread of this virus in India.

Along with visualizing the above mentioned counts we have also analyzed available data of the patients to find if age bracket has any relation with the spread of this virus. Our visualizations also reflect how imported cases played a major role in initiating the outbreak in India.

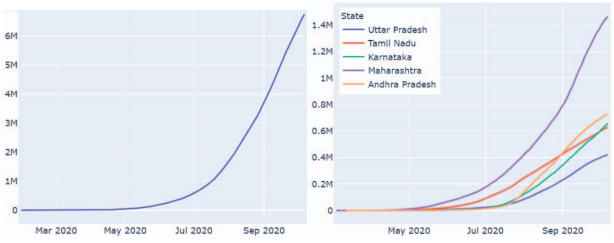
A. Confirmed Cases

The count of people C_t that has got infected from the virus on or before a particular day t is the count of *Confirmed Cases* until that day. This quantity is a cumulative value because it doesn't reduce when the patient gets cured. This statistic simply records the count of people who got exposed to the virus and got infected.

Hence, it can be stated that count of confirmed cases C_{t+1} on day $t+1$ is always greater than or equal to the confirmed cases C_t on day t.

$$C_{t+1} >= C_t \tag{1}$$

From Fig. 1a, it can be seen that the number of confirmed cases has been increasing over the months and there has been a significant rise in the number of cases from July 2020. Similar trends can be observed for the spread of COVID-19 in individual states, where Uttar Pradesh has the maximum number of confirmed cases in India, as shown in Fig. 1b.



(a) Spread in the country (b) Spread in the states

Fig. 1: Rise in number of cases with time (Count in Millions)

B. New Cases

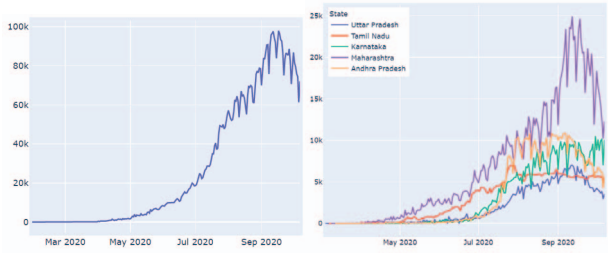
The number of people N_t that are tested positive for the virus on a given day t is the count of *New Cases* on that day. Hence, if C_{t-1} is the number of confirmed cases on $t-1_{th}$ day and C_t is the number of confirmed cases on t_{th} day then

$$N_t = C_t - C_{t-1} \tag{2}$$

This quantity is not cumulative. The count of new cases N_t on day t can be less than, equal to or even greater than count

of new cases N_{t+1} on day $t+1$. Thus we can state that N_t has no *mathematical* relation with N_{t+1}.

Fig. 2a shows the analysis of new cases observed in the country. Trends indicate massive growth in the initial months and depreciation in late September and October. The spread of new cases in different states can be seen in Fig. 2b. Even here a similar trend can be visualized. Trends of initial months have witnessed a steep growth in daily new cases which has slowed down and even started falling in the later half.



(a) Spread in the Country (b) Spread in the states

Fig. 2: New COVID-19 Cases over time

According to the *SIR Epidemic Model*, [19], in the beginning, the entire population is at the risk of getting infected by the virus. Hence, theoretically, when the entire population has got infected from the virus, the count of new cases from there onward will become 0, given that a person cannot get infected from the virus again, once he/she develop immunity against it.

Hence,

$$N_\infty = 0 \tag{3}$$

C. Recovered Cases

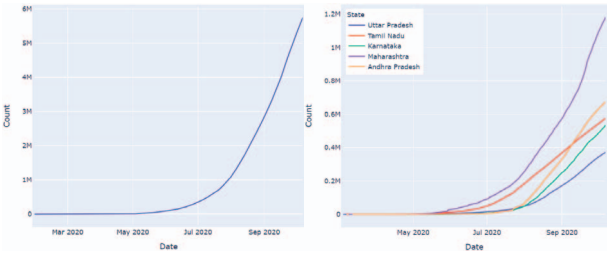
The count of people R_t, who have been detected positive and eventually got cured, on or before a particular day t is the count of *Recovered Cases* till that day. Even this quantity is cumulative. As an infected patient gets cured, the count keeps increasing. Hence,

$$R_{t+1} >= R_t \tag{4}$$

Fig. 3a shows the history of the cumulative count of recovered cases in India with time. The exponentially rising graph indicates a high recovery rate (Section III-G) in India. It can be gained from the figure that close to 6 million patients of COVID-19 in India have recovered from the infection. Fig. 3b shows a history of recovered cases in the adopted states. Even though these states are the worst affected states, it's a good thing to note that the rise in recoveries in these states is also high as evident from the figure.

D. Deaths

The total number of people D_t who were detected positive and died due to the virus on or before a particular day t is the count of *Deaths* till that day. Fig. 4a and 4b show the

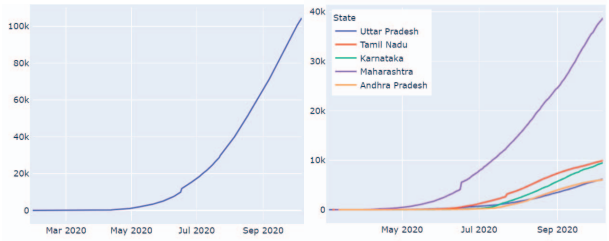


(a) Recovery in the country (b) Recovery in the states

Fig. 3: Recovery of cases in India and states with time

rise in the number of deaths with time nationwide and state-wise, respectively. It can be inferred that even this quantity is cumulative and the number of deaths has been rising as observed from February to October 2020. Hence,

$$D_{t+1} >= D_t \tag{5}$$



(a) Deaths in the country (b) Deaths in the states

Fig. 4: Rise in Deaths due to COVID-19 with time

E. Active Cases

The number of *Active Cases* I_t on a particular day t is the count of people who are currently infected from the virus and not cured.

Any person that is tested positive leads to an increment in the count of confirmed as well as active cases. This active case becomes inactive either when the patient gets cured or if the patient dies in the struggle to get recovered.

This leads us to the conclusion that at a given point of time t, a confirmed case is either an active case, a recovered case or a death case. Hence we can say that,

$$C_t = I_t + R_t + D_t \tag{6}$$

Hence at any point of time t, number of active cases I_t is the subtraction of recoveries R_t and deaths D_t from total number of confirmed cases C_t.

$$I_t = C_t - (R_t + D_t) \tag{7}$$

The number of active cases in the country has been increasing tremendously since the inception of COVID-19 in India,

but a declining pattern can be observed from September 2020 as in Fig. 5a. The patterns of the number of active cases state-wise vary from state to state as in Fig. 5b, where most of the states show a decrease in the number of active cases from September 2020.

Similar to the count of new cases, even active cases eventually become 0. Hypothetically, one day ct, when the entire population is already affected by the virus, i.e. when C_{ct} is equal to count of the population, there will be no more new cases. Hence N_{ct+1} will be 0 since from the next day onward there will be no susceptible person in the population. But this in no way asserts that there will be no active cases. All we know is that from $ct + 1^{th}$ day we are not having any new cases.

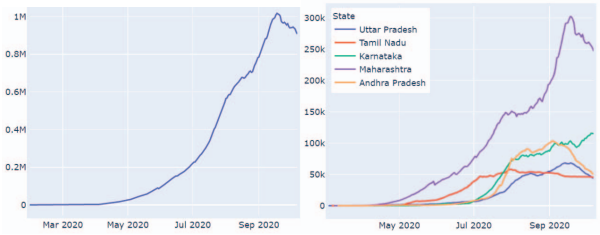
Now, let us assume that the last patient to remain infected by the virus gets cured or dies, δ days after ct. This means that we have the last active case on $ct + \delta^{th}$ day. Hence,

$$I_{ct+\delta} > 0 \tag{8}$$

and according to our hypothesis, there are no new susceptible people after ct^{th} day, hence from $ct + \delta + 1^{th}$ day onward there won't be any active cases.

$$I_{ct+\delta+1},\ I_{ct+\delta+2,\ldots} = 0 \tag{9}$$

where ct is the day when the last person in the population gets infected by the virus and δ is the number of days after ct, when the last person gets cured of the disease.



(a) Active Cases in the country (b) Active Cases in the states

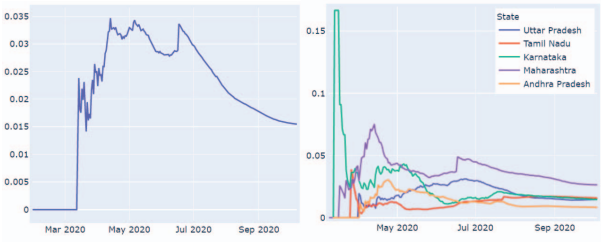
Fig. 5: Active COVID-19 Cases

F. Mortality Rate

In the context of COVID-19, Mortality Rate or Death Rate is the fraction of deaths due to COVID-19 over the total count of infected patients. Hence, on a given day t,

$$Mortality\ Rate = D_t/C_t \tag{10}$$

Fig. 6a and 6b show the mortality rate in India and the state-wise numbers, respectively. It is clear that during the initial months of the outbreak in the country, the death rate has been increasing, which could be because of the unpreparedness towards the situation. But eventually, there has been a steady decline in the mortality rate.



(a) Mortality Rate in the country (b) Mortality Rate in the states

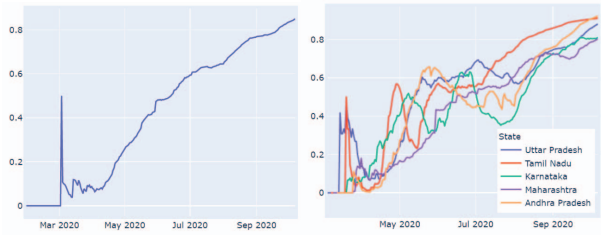
Fig. 6: Mortality Rate in India against COVID-19

G. Recovery Rate

Recovery Rate is the fraction of recovered cases over the total count of infected patients. Hence, on a given day t,

$$Recovery\ Rate = R_t/C_t \tag{11}$$

In the beginning, the recovery rate against COVID-19 in India was quite less which could again be because of the sudden surge in the number of cases and lack of facilities, but eventually, the recovery rate has been growing, as it is distinctly observable in Fig. 7a (for the overall country) and 7b (for different states).



(a) Recovery Rate in the country (b) Recovery Rate in the states

Fig. 7: Recovery Rate in India against COVID-19

H. Daily Infection Rate

Daily Infection Rate (DIR) is defined as the rate of change in active cases with respect to the previous day.

$$Daily\ Infection\ Rate\ = \frac{I_t - I_{t-1}}{I_{t-1}} \tag{12}$$

where I_t and I_{t-1} are the number of active cases on day t and $t-1$ respectively. The DIR takes a positive value when there is a rise in an active case, becomes 0 when the number of active cases does not change and becomes negative when the number of active cases decreases with respect to the previous day. Fig. 8 shows DIR of India from the initial months of COVID-19 in India up to 7th October 2020. It can be seen that DIR has been dropping right from the beginning and in late September and October, the rate has become negative which indicates a drop in active cases.

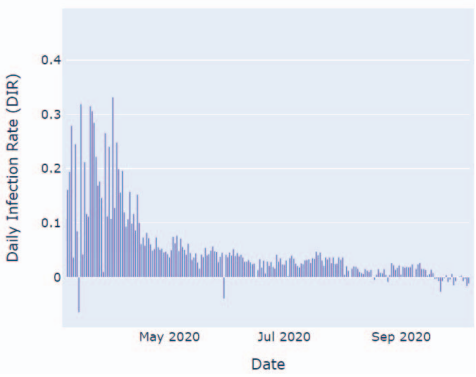


Fig. 8: Daily Infection Rate of COVID-19 in India

IV. Time Series Forecasting Algorithms

Time Series forecasting algorithms comprise a class of Machine Learning algorithms that focus on understanding and approximating behavior, trend and other properties of sequential data.

For forecasting the COVID-19 outbreak in India we have specifically used the *ARIMA Forecasting Model* and the *Prophet Model by Facebook*.

Using these two algorithms we have predicted the spread (confirmed cases), deaths, recoveries and active cases in India as well as in the states that have been adopted for our research. Our study also involves a comparison of outputs generated by these two models for different time series with the help of relevant performance metrics (described in Section V-B). An overview of the above-mentioned algorithms is given in the following sections.

A. Autoregressive Integrated Moving Average (ARIMA)

ARIMA(p,d,q) [20] is one of the most renowned and widely used statistical models for time series forecasting. ARIMA is a composite model, the key aspects of which lie right in its name. These components are Autoregression (AR), Integrated (I) and Moving Average (MA).

The parameters p, d and q control the above-mentioned aspects of this model respectively. Hence, the correct notation to describe an ARIMA model is ARIMA(p,d,q) where these parameters are substituted with integer values to indicate the specific ARIMA model that is being used.

The Autoregressive AR(p) model is an autoregression model that tries to estimate and model the relationship between an observation and its lagged observations. The number of lagged observations that this model considers, depends on the value of p. The AR(p) model is defined by the following equation:

$$z_t = c + \sum_{i=1}^{p} \phi_i z_{t-i} + \varepsilon_t \tag{13}$$

where $z_{t-1},\ z_{t-2},\ z_{t-3}, ...,\ z_{t-p}$ are the lagged observations; $\phi_1,\ \phi_2,\ \phi_3, ...,\ \phi_p$ are the coefficients that are approximated by the model to fit onto the data; ε_t is the white noise and c is a constant.

The Moving Average MA(q) model is defined by the following equation:

$$z_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{14}$$

where $\theta_1,\ \theta_2, ...,\ \theta_q$ are parameters of the model and μ is the expectation of z_t which is often assumed to be equal to 0 and $\varepsilon_t,\ \varepsilon_{t-1}, ...,\ \varepsilon_{t-q}$ are white noise error terms for the observations at $t,\ t-1, ...,\ t-q$ respectively.

The AR(p) model and MA(q) model together form the Auto Regressive Moving Average (ARMA) model [21] [22]. The problem with ARMA is that they are suited to work with stationary time series only. Stationary time series are time series with a constant mean and variance, hence having no trend or seasonality.

ARIMA(p,d,q) is different from ARMA(p,q) due to the "integrated" component. This component solves the major drawback of the ARMA model, which is the model's inability to effectively fit onto non-stationary time series data. ARIMA solves this drawback by introducing differencing that is controlled by a new parameter d, which is the differencing parameter. Differencing in statistics is a transformation applied to time series data to make it stationary. The value of parameter d states the order of differencing. First-order differencing ($d = 1$) is done by calculating the difference between consecutive observations.

First order differencing ($d = 1$) is represented as:

$$z'_t = z_t - z_{t-1} \tag{15}$$

Similarly, second order differencing ($d = 2$) is represented as:

$$z''_t = z'_t - z'_{t-1} \tag{16}$$

Here z'_t and z''_t are first and second-order differences for z_t. Hence an ARIMA(p,d,q) model initially performs differencing on the time series as per d. This differencing when done effectively removes non-stationarity in the time series. Then the model fits the generated stationary time series in the following ARIMA model equation which is simply a linear combination of (13) and (14).

$$z_t = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i z_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{17}$$

B. Prophet by Facebook

Taylor et al. [23] proposed the Facebook Prophet (FBProphet) which uses several methods as components for time series forecasting. Prophet is developed and released as open-source software by the data science team of Facebook. Prophet facilitates forecasting of time-series using simple intuitive parameters and supports including the effect of seasonality and holidays as well.

Prophet model has three main components: trend, seasonality and holidays, which can be expressed using the following equation:

$$y_t = g_t + s_t + h_t + \varepsilon_t \tag{18}$$

where g_t is trend; s_t is seasonality; h_t is holiday and ε_t is error term.

V. Experiments

A. Dataset

The dataset [18] that has been adopted for our research is an open dataset on Kaggle. The dataset has records of tests being conducted in different states and union territories of India every day and also the count of positive and negative cases out of the total tests conducted. It also records deaths, confirmed cases and cured cases on a day-to-day basis. This information has been provided in a state-wise manner. Hence, preprocessing is required before any nationwide analysis can be performed using the data.

B. Performance Metrics

For the evaluation of the forecasting models, we have used the following statistical metrics.

1) Root Mean Square Error (RMSE): Root Mean Squared Error is used to measure the mean of the difference in the values predicted by the model and the observed values.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y_i})^2}{n}} \tag{19}$$

2) Mean Absolute Error (MAE): Mean Absolute Error is defined as the average of the absolute errors between the predicted and the observed values.

$$Absolute\ Error = |y_i - \hat{y_i}| \tag{20}$$

$$MAE = (\frac{1}{n}) \sum_{i=1}^{n} |y_i - \hat{y_i}| \tag{21}$$

3) Mean Absolute Percentage Error (MAPE): Mean Absolute Percentage Error [25] measures the accuracy of the forecasting model in terms of the average of the percentage errors between the actual and the predicted values.

$$MAPE = (\frac{100}{n}) \sum_{i=1}^{n} \frac{|y_i - \hat{y_i}|}{y_i} \tag{22}$$

4) Coefficient of Determination (R^2): Coefficient of Determination (R^2) [24] is the measure of how close the predicted values are to the line of regression.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y_i})^2}{\sum (y_i - \overline{y})^2} \tag{23}$$

where y_i is the ground truth value at i_{th} observation; $\hat{y_i}$ is the predicted value at i_{th} observation based on the hypothesis of the predicting model; $\overline{y}$ is the mean of all observations y_i; and n is the total number of observations.

VI. Results

We have trained the above discussed models (Section IV) to forecast confirmed cases, death cases and recoveries in India as well as in the adopted states. Each of these models is trained and tuned on more than 180 days of data, starting right from the day on which the above regions reported a confirmed case of COVID-19 for the first time up till 8th September 2020. To identify appropriate values of p, d and q for ARIMA models we have used a grid search approach. The orders of ARIMA models are mentioned in the tables below. On the other hand, Prophet model was directly applied to the actual data. The tuned models are then evaluated using appropriate performance metrics (Section V-B) over 28 days of data, i.e. from 9th September 2020 to 7th October 2020. After validation, tuning and evaluation of these models, we use them to forecast counts for the next 14 days i.e. from 8th October 2020 to 21st October 2020. For this forecast of 14 days, the models are not evaluated. This is for simply comparing the forecasts of the two models visually.

For comparative purposes, forecasting results for the adopted states are studied together and results at the country-level i.e. for India are studied separately. Firstly we will compare the results that we have achieved at the state level.

A. Forecasting results for confirmed cases

Table I shows metric scores of ARIMA and Prophet Model for confirmed cases among the adopted states. Best forecasting results have been achieved for the state of Tamil Nadu with R^2 scores 0.9984 and 0.9869 for ARIMA and Prophet respectively. Minimum and maximum R^2 values of ARIMA are 0.6801 and 0.9984 respectively. On the other hand minimum and maximum R^2 values of Prophet model are -0.0966 and 0.9869 respectively.

TABLE I: Metric results for forecasting of confirmed count of cases in the adopted states *(trained on more than 180 days of data, till September 9, 2020)*

State	Model	RMSE	MAE	MAPE	R^2
Maharashtra	ARIMA(2,1,1)	20814.98	18879.21	1.542	0.9811
	Prophet	158559.55	151146.35	11.866	-0.0966
Tamil Nadu	ARIMA(2,1,1)	1764.63	1513.16	0.261	0.9984
	Prophet	5137.09	4372.37	0.752	0.9869
, Andhra Pradesh	ARIMA(10,1,1)	33915.84	25718.67	3.750	0.6801
	Prophet	21948.27	16304.62	2.404	0.8660
Uttar Pradesh	ARIMA(5,1,1)	14085.58	9368.95	2.353	0.8802
	Prophet	21338.40	20853.88	5.786	0.7251
Karnataka	ARIMA(5,1,1)	13478.95	12130.94	2.153	0.9620
	Prophet	28396.77	26978.43	4.878	0.8314

B. Forecasting results for death cases

Table II shows metric scores of ARIMA and Prophet Model for death cases among the adopted states. We can see that the best forecasting results have been achieved for the state of Karnataka with R^2 scores 0.6697 and 0.9665 for ARIMA

and Prophet respectively. Minimum and maximum R^2 values of ARIMA are 0.6697 and 0.9924 respectively. On the other hand minimum and maximum R^2 values of Prophet model are -1.1486 and 0.9665 respectively.

TABLE II: Metric results for forecasting of deaths cases in the adopted states *(trained on more than 180 days of data, till September 9, 2020)*

State	Model	RMSE	MAE	MAPE	R^2
Maharashtra	ARIMA(5,1,2)	1500.21	1343.25	3.844	0.7949
	Prophet	1872.59	1704.87	4.901	0.6805
Tamil Nadu	ARIMA(10,1,1)	263.83	234.10	2.529	0.7674
	Prophet	802.03	735.13	7.981	-1.1486
, Andhra Pradesh	ARIMA(10,1,1)	36.87	31.15	0.561	0.9924
	Prophet	524.98	446.18	7.877	-0.5246
Uttar Pradesh	ARIMA(5,1,1)	164.39	145.12	2.661	0.9308
	Prophet	346.82	323.03	6.006	0.6924
Karnataka	ARIMA(10,1,1)	446.56	361.98	4.162	0.6697
	Prophet	142.12	127.90	1.538	0.9665

C. Forecasting results for recovered cases

Table III shows metric scores of ARIMA and Prophet Model for recoveries among the adopted states. The best forecasting results have been achieved for the state of Tamil Nadu with R^2 scores 0.9638 and 0.9957 for ARIMA and Prophet respectively. Minimum and maximum R^2 values of ARIMA are 0.5740 and 0.9938 respectively. On the other hand minimum and maximum R^2 values of Prophet model are 0.1995 and 0.9957 respectively.

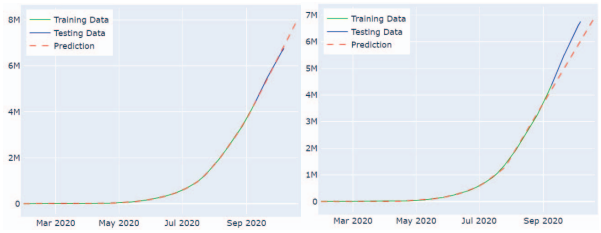
TABLE III: Metric results for forecasting of recoveries in the adopted states *(trained on more than 180 days of data, till September 9, 2020)*

State	Model	RMSE	MAE	MAPE	R^2
Maharashtra	ARIMA(1,1,2)	102654.98	84119.39	8.155	0.5740
	Prophet	140734.66	120334.20	11.875	0.1995
Tamil Nadu	ARIMA(2,1,1)	8565.99	6844.23	1.286	0.9638
	Prophet	2944.11	2568.76	0.502	0.9957
, Andhra Pradesh	ARIMA(2,1,1)	25447.27	19494.73	3.193	0.8856
	Prophet	25766.39	24961.01	4.447	0.8827
Uttar Pradesh	ARIMA(3,1,1)	9521.37	8242.43	2.590	0.9598
	Prophet	33356.48	30583.24	9.818	0.5067
Karnataka	ARIMA(1,1,1)	4997.74	3816.77	0.932	0.9938
	Prophet	28968.13	27441.14	6.228	0.7924

D. Forecasting results for India

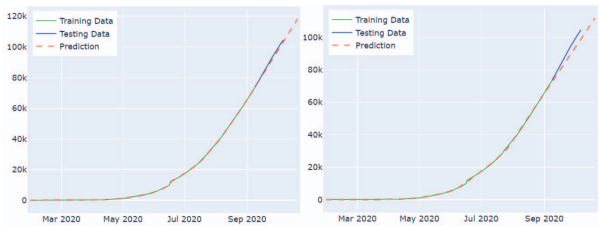
ARIMA and Prophet forecasting models are trained and tuned on precisely 224 days of data from 30th January 2020 to 8th September 2020 for forecasting the COVID-19 spread in India. The metric scores that can be seen in Table IV are calculated by evaluating the predictions of trained models on 28 days of data (9th September to 7th October 2020) against ground truth values. We have also put these models into action by forecasting counts for the next 14 days (8th October to 21st October 2020). The forecasts for these 14 days are not tested against any ground truth values, rather visualized.

The achieved results (as mentioned in Table IV) clearly indicate that ARIMA has outperformed Prophet Model in forecasting future COVID-19 counts. Fig. 9a, 10a and 11a show predictions of ARIMA model for confirmed cases, death counts and recoveries in India respectively. On the other hand Fig. 9b, 10b and 11b show forecasts of Prophet model for the same. It is quite evident from the figures as well that ARIMA model performance is better than that of Prophet model. But this does not imply that Prophet model needs to be ruled out. The metric scores of the model are still acceptable.



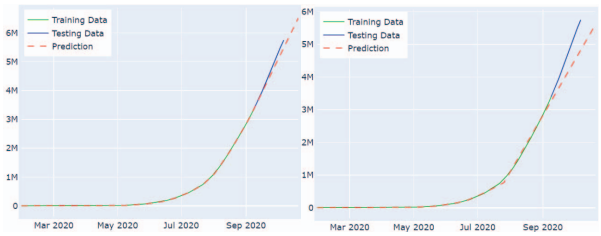
(a) ARIMA(1,1,1) forecasting (b) Prophet forecasting

Fig. 9: Forecasting confirmed cases of COVID-19 in India



(a) ARIMA(1,1,1) forecasting (b) Prophet forecasting

Fig. 10: Forecasting COVID-19 related deaths in India



(a) ARIMA(3,1,1) forecasting (b) Prophet forecasting

Fig. 11: Forecasting recovered cases of COVID-19 in India

TABLE IV: Metric results for forecasting of confirmed cases, deaths and recoveries in India *(trained on precisely 224 days of data, till September 9, 2020)*

State	Model	RMSE	MAE	MAPE	R^2
Confirmed Cases	ARIMA(1,1,1)	39275.41	33615.02	0.592	0.9967
	Prophet	544986.48	519091.46	8.928	0.3749
Deaths	ARIMA(1,1,1)	1180.22	1097.33	1.184	0.9823
	Prophet	4324.34	4091.15	4.417	0.7632
Recovered Cases	ARIMA(3,1,1)	168807.59	137610.35	2.722	0.9412
	Prophet	596137.98	541964.12	11.180	0.2674

VII. Conclusion

In this paper, we have done the data analysis, visualization and prediction study of the COVID-19 pandemic outbreak in India. Data analysis and visualization studies have been done on more than 8 months of data. For the forecasting study, we have used two of the most widely used time-series forecasting models; ARIMA and Prophet. Rather than just localizing our research towards India, we have also done a comparative state-wise study among the 5 worst-affected states of India. The forecasting results have been quantified and compared using performance metrics like RMSE, MAE, MAPE and R^2. Achieved scores indicate that ARIMA has outperformed Prophet in forecasting the counts. The trend analysis shows rapid growth in the infected cases and the prediction study shows a huge rise in confirmed cases, recoveries and death cases in India. However, lockdowns and containment policies may affect the prediction. India has lost more than 100,000 lives due to COVID-19 and the count is still increasing. On the contrary, good thing is that the recovery rate in India is quite high and the daily infection rate (DIR) is also dropping right from the initial months of spread itself. Future work in forecasting the COVID-19 pandemic includes taking into account various other factors as well, such as population density, weather, health system, patient history, etc. Ensemble methods and Deep Learning approaches can also help in achieving better forecasting results.

References

[1] Coronavirus disease: What you need to know. World Health Organization. https://www.afro.who.int/news/coronavirus-disease-what-you-need-know.

[2] Ghosh P, Ghosh R, Chakraborty B. "COVID-19 in India: Statewise Analysis and Prediction," JMIR Public Health Surveill. 2020;6(3):e20341. Published 2020 Aug 12. doi:10.2196/20341

[3] Lin Jia, Kewen Li, Yu Jiang, Xin Guo, et al. "Prediction and analysis of coronavirus disease 2019," arXiv preprint arXiv:2003.05447, 2020.

[4] N. Kumar and S. Susan, "COVID-19 Pandemic Prediction using Time Series Forecasting Models," *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020, pp. 1-7, DOI: 10.1109/ICCCNT49239.2020.9225319.

[5] Dehesh, T., Mardani Fard, H. A., Dehesh, P. (2020). "Forecasting of COVID 19 Confirmed Cases in Different Countries with ARIMA Models," *medRxiv*.

[6] Shi, Z., Fang, Y. (2020). "Temporal relationship between outbound traffic from Wuhan and the 2019 coronavirus disease (COVID 19) incidence in China."

[7] Rajan Gupta, Saibal Kumar Pal (2020)."Trend Analysis and Forecasting of COVID-19 outbreak in India," *medRxiv*, 2020.03.26.20044511; doi: https://doi.org/10.1101/2020.03.26.20044511

[8] F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in IEEE Access, vol. 8, pp. 101489-101499, 2020, doi: 10.1109/ACCESS.2020.2997311.

[9] S. S. Arun and G. Neelakanta Iyer, "On the Analysis of COVID19 - Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2020, pp. 1222-1227, doi: 10.1109/ICICCS48265.2020.9121027.

[10] Date, Saroj and Deshmukh, Sachin. (2020). "Forecasting novel COVID-19 confirmed cases in India using Machine Learning Methods," *INTERNATIONAL JOURNAL OF COMPUTER SCIENCES and ENGINEERING*. 8. 57-62. 10.26438/ijcse/v8i6.5762.

[11] M. Mousavi, R. Salgotra, D. Holloway and A. H. Gandomi, "COVID-19 Time Series Forecast Using Transmission Rate and Meteorological Parameters as Features," in *IEEE Computational Intelligence Magazine*, vol. 15, no. 4, pp. 34-50, Nov. 2020, doi: 10.1109/MCI.2020.3019895.

[12] Chimmula, Vinay Kumar Reddy and Lei Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, solitons and fractals* vol. 135 (2020): 109864. doi:10.1016/j.chaos.2020.109864

[13] Bouhamed, Heni. (2020). "Covid-19 cases and recovery previsions with Deep Learning nested sequence prediction models with Long Short-Term Memory (LSTM) architecture." 8. 10-15.

[14] V. Z. Marmarelis, "Predictive Modeling of Covid-19 Data in the US: Adaptive Phase-Space Approach," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 207-213, 2020, doi: 10.1109/OJEMB.2020.3008313.

[15] S. Singh, P. Raj, R. Kumar and R. Chaujar, "Prediction and forecast for COVID-19 Outbreak in India based on Enhanced Epidemiological Models," *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India, 2020, pp. 93-97, doi: 10.1109/ICIRCA48905.2020.9183126.

[16] Anastassopoulou Cleo, Russo Lucia, Tsakris Athanasios and Siettos Constantinos, "Data-based analysis modeling and forecasting of the COVID-19 outbreak," Plosone, March 2020.

[17] A. Bhati and A. Jagetiya, "Prediction of COVID-19 Outbreak in India adopting Bhilwara Model of Containment," 2020 5th International Conference on Communication and Electronics Systems (ICCES), COIMBATORE, India, 2020, pp. 951-956, doi: 10.1109/ICCES48766.2020.9138060.

[18] Sudalai Rajkumar. (2020, March). "COVID-19 in India,". Retrieved October 20, 2020 from https://www.kaggle.com/sudalairajkumar/covid19-in-india.

[19] Kermack William Ogilvy and McKendrick A. G. 1927, "A contribution to the mathematical theory of epidemics," Proc. R. Soc. Lond. A115700–721.

[20] Box, George E. P., Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung. 2016."Time series analysis: forecasting and control".

[21] Hannan, Edward James (1970). "Multiple time series," *Wiley series in probability and mathematical statistics.* New York: John Wiley and Sons.

[22] Whittle, P. (1951). "Hypothesis Testing in Time Series Analysis," Almquist and Wicksell. Whittle, P. (1963). *Prediction and Regulation*. English Universities Press.

[23] Sean J Taylor and Benjamin Letham. "Forecasting at scale," *The American Statistician*, 72(1):37–45, 2018

[24] Wright, Sewall. 1921. "Correlation and causation," *Journal of Agricultural Research* 20: 557-585.

[25] Willmott, Cort J.; Matsuura, Kenji (December 19, 2005). "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". *Climate Research.* 30: 79–82.