# Predicting COVID-19 Cases in India using Global Data

**Shubhnesh Kumar Goyal**

Associate Professor, D.S. (P.G.) College, Aligarh, India

**Abstract:** ***This study presents a novel method of using SIR-based epidemiological modeling of COVID-19 spread to predict the number of cases in India during January 2020 – April 2020 by identifying and leveraging the information from other countries that evolved through the same phase. Firstly, countries with similar characteristics were identified using R2 scores, and then their parabolic parameters were used to predict the number of cases in India 15-30 days in advance. Results show excellent accuracy for mentioned period, and thus this method put forward a promising way to mitigate upcoming damage by proper planning.***

**Keywords:** Mathematical modeling, Epidemiology, SIR models, Statistical prediction, Optimisation

## 1. Introduction

India remains one of the most vulnerable targets of the COVID-19 pandemic due to a large population base. The government has mitigated almost every department to fight this global problem and there have been pressing demands of forecasting of confirmed cases so that proper steps can be taken beforehand to deal with the disaster we are facing. With the advent of global pandemic, we first had to deal with 2-3 cases per day coming mainly from international territories and these were well quarantined. But as the cases increased so is their rate of increasing per day, and this problem is generally explained using exponential trends due to multiple hosts. A brief analysis of the healthcare system, help us to conclude that we have a fairly strong system that can deal with small number of cases very well, but as the cases will rise, so is the number of patients needing intensive care, and this will soon lead to health bed limitation as well as health care equipment and official's saturation. In such a scenario, having an estimate of how things will progress, what is the preparation we should have and when will the worst strike us can really help.

Best way to tackle this problem is to study epidemiology of this specific virus and then having extensive simulations specific to regions with varied demographic. Though this process will yield accurate results, this will take time to develop and till then improved simple statistical models can provide a safety net to fall on. There are mainly four categories into which these types of modelling can be grouped:

1) Simple exponential trends, and logistic regression-based estimates
2) Deep Learning based approach with fine hyperparameter tuning and PCA
3) SIR modelling with improved parameters like localised R-rate and SIER models
4) Simulation based models with individual interaction modelled over networks

All of these methods tend to capture the epidemiological properties and progress trend of virus and the population. This paper is an attempt to solve the problem using novel statistical approach based on SIR model. While Deep Learning has been suggested for near days forecasting, this method focuses on the overall dynamics of the disease till saturation. Detailed plots and analysis are given for work done and the work can be considered as exploratory research work on which the foundation of better models can be established.

## 2. Dataset Preparation

International data was taken from John Hopkins maintained database [5, 4, 6] and national data was taken from the government-maintained dataset on the COVID-19 Tracker website. [7, 8] MinMax scaling of the data is done in all of the models and the scaler is fitted only on the training data to prevent the forward bias. Two different sets of features are also used with details given afterward. Data till 28th April (from 14th March for India) was used for training and testing purposes, while the data till 29th April was available. Other minor changes in the data are mentioned along with the method.

### Data Visualisation

Different aspects of data can be visualized through time series plots and we show two of them mainly, total number of cases confirmed each day cumulatively (fig 1.1) and daily difference for India (fig 1.2).

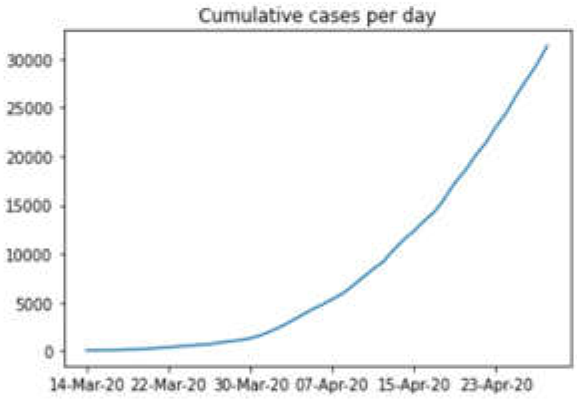


**Figure 1.1**

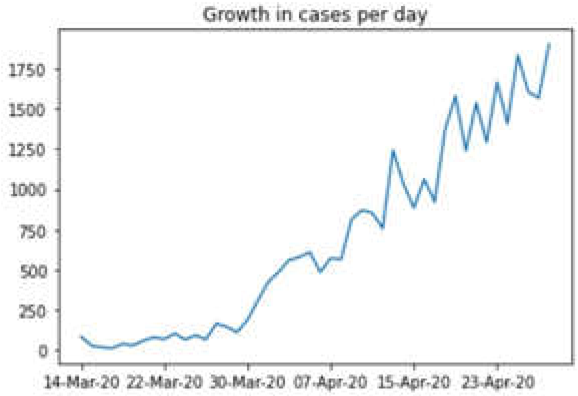


**Figure 1.2**

### SIR modeling:

SIR model is a simple statistical model which governs the growth of total number of cases of an infectious disease. The model consists of three group of people, namely susceptible people, infectious people and recovered people. Susceptible group starts from the whole population, and only way to leave susceptible is to get infected. For this a parameter is introduced say b, which is the number of people one infectious agent will meet and transfer the disease every day. Similarly recovered people are calculated using a recovery time period say k, so any day, 1/k people will recover. These results in three differential equations [10] which when solved, give dynamics of disease increase.

$$\frac{ds}{dt} = -b\, s(t)\, i(t),$$

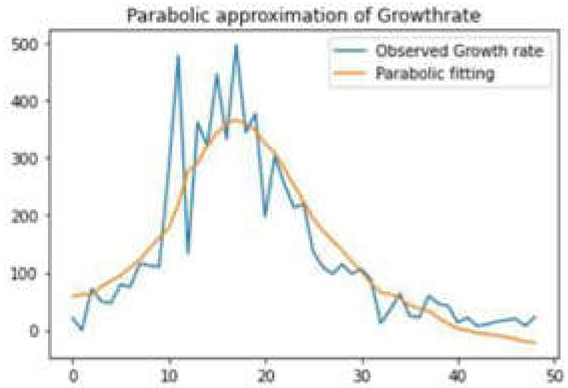
$$\frac{di}{dt} = b\, s(t)\, i(t) - k\, i(t),$$

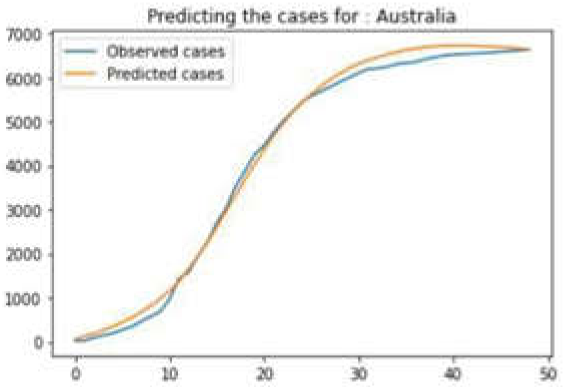
$$\frac{dr}{dt} = k\, i(t).$$

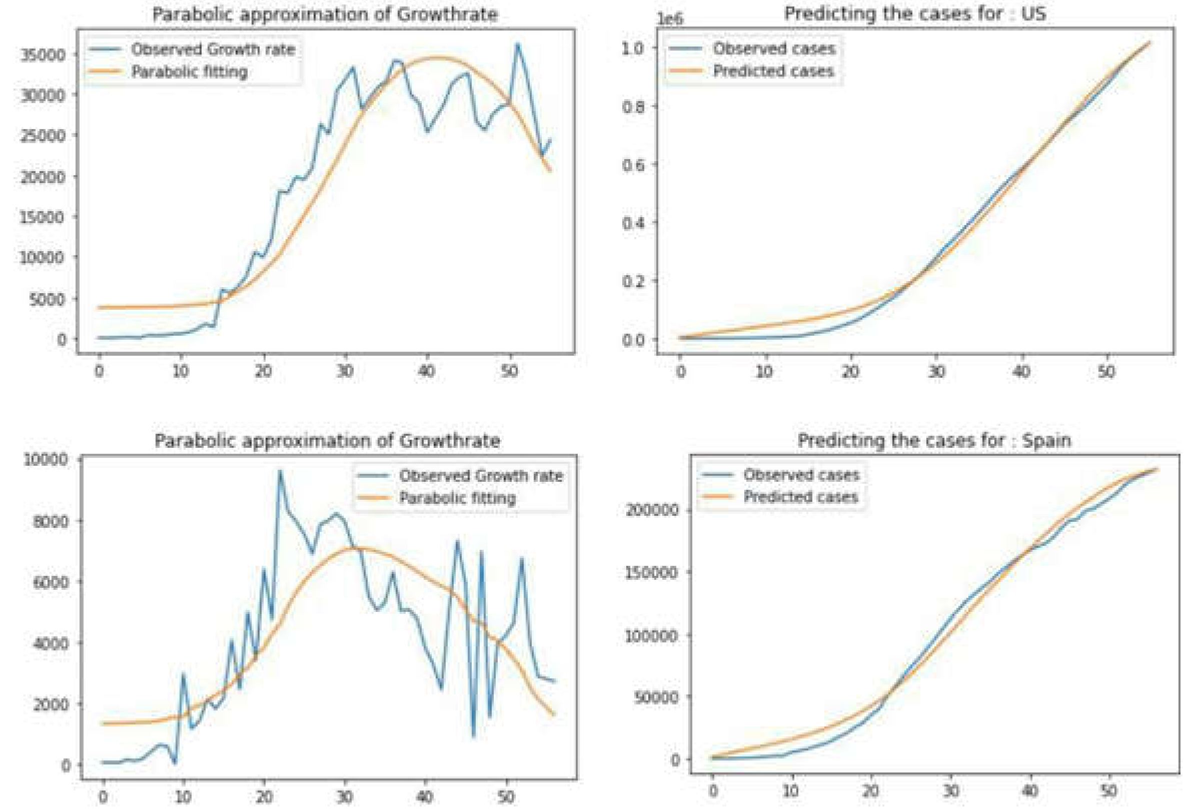
But the problem remains in finding the said introduced parameters, specifically the b, since it largely depends on administrative steps, population demographics and quarantine condition. Also, k is dependent on virus, temperature, healthcare facility and many other factors. These limitations present a roadblock in using this model in early stages of any disease, since a large amount of data is needed to calculate k and b.

### Parabolic analysis based on the SIR model:

Although Deep Learning based models provide a good way of approximating near future values, all of them **lack the ability to predict the general trend** which the disease dynamics will follow. To capture this, here we present a way which kind of **learns from how the other countries behaved during the phase we are in right now and how that ended up for them**.







Any infectious disease which follows the SIR model, its **growth rate vs cumulative cases can be modelled as a parabola [3]**. When we applied this to the countries that are in **late phases of the epidemic like Australia**, for countries in the **third phase, like America and Spain** and for **India** excellent results (fig 2) were found.

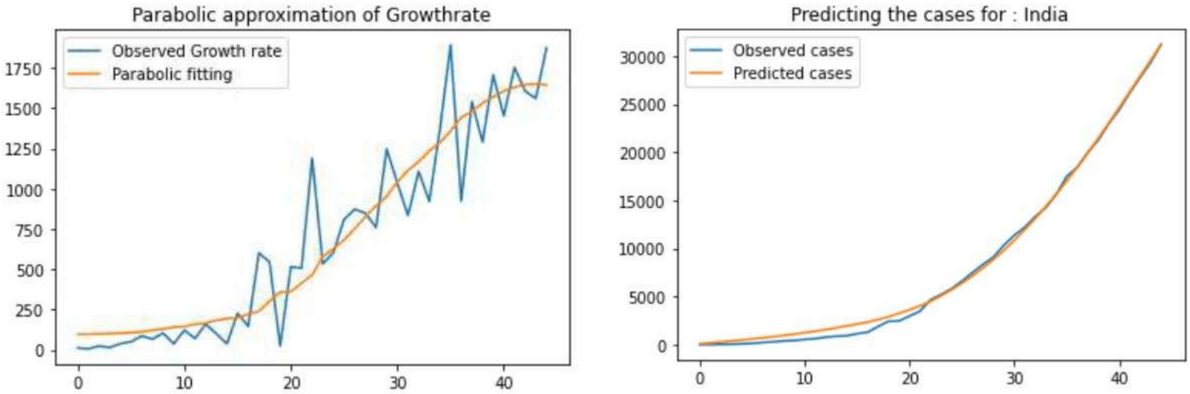


**Figure 2:** Represents the fitting of parabola on growth rate (y-axis) vs total cumulative days on particular days (x-axis) for four different countries, and prediction using this for cumulative case (y-axis) vs number of days (x-axis)

This could also be used for prediction of the country's cases by following the similar trend forward but this presents one problem, **if the points lie on only one arm of parabola, then, while fitting last point is assumed to be the almost maximum point and further predictions come with a downward slope (fig 3)**. If the points lie on both the sides of approximated parabola, then future predictions come in vicinity of actual observations.

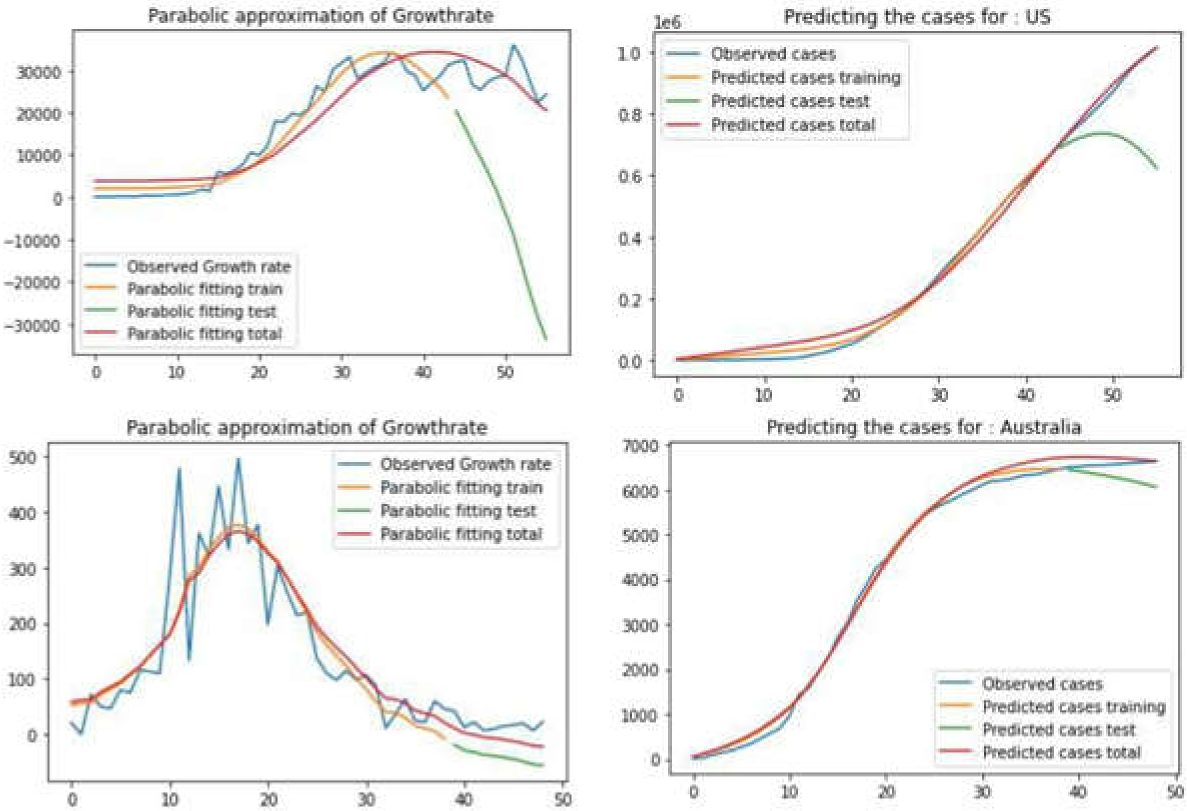


**Figure 3:** Represents the predicted cases using 80% of the data points vs predicted cases using 100% of the data (y-axis) plotted along number of days (x-axis) for two different countries.

Though this model is not good in direct predictions by training on some data, we present another **novel approach of predictions using another country's data**.

We did a cross correlation test, in which t**he model was trained on one country** and it was **tested on another country** going through different phases. As though the model learned different behaviours of the virus evolution, this yielded good results. **Models trained on similar countries yielded excellent predictions for another, and even the deviation as the epidemic progressed gave way to in depth analysis.** To better explain the given fact, we present a case of **Italy and Spain (till 14th April)** (fig 4)

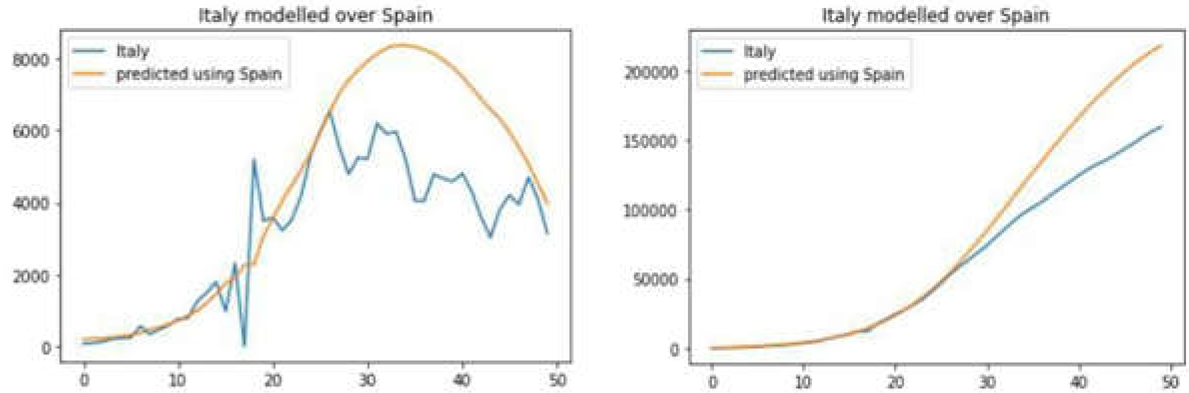


**Figure 4:** Represents predicting number of observed cases in Italy (y-axis) using only the Spain's data

One key point here is that the parabola which is used to predict Italy has only seen Spain's data. Still, it fits in the starting period or 25 days till 50,000 cases. This shows that till then **both of the countries were evolving under similar dynamics** but if Italy were to follow a similar course, then it should end up at .2 million cases, but it would be saturated within much lower values after deviation around the 27th day.

After some analysis of the two countries, it was found that both of them have almost similar health workforce, but in the **middle of the crisis due to less availability of PPEs in Spain, its 21% workforce got affected by the virus, while that number stayed below 10% for Italy [9]**.

This prompted us to find those countries that were in **similar condition as India some time during their course of evolution**, and if we can find those, then we can find where India would end up according to each country's dynamics. Then we can have a guideline on how to act accordingly and is presented below,

First step was to find countries that were having similar dynamics as us in this phase, and for that we **trained a different model on each country till 14th April** that has more cases than us and then **tested this model on our country till 14th April**. We chose all the countries that had a **R2 score of greater than 0.9** on predicting India's trend.

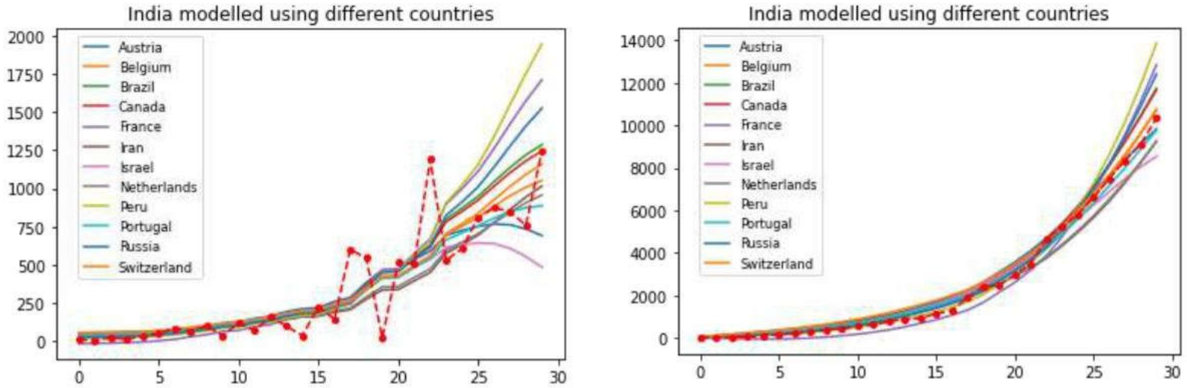


**Figure 5:** Represents predicting observed cases (x-axis) in India using the data from other countries.

This shows the similarity between India's growth in this period and how other countries fared when they were in similar numbers. Now, each curve here represents how that country evolved and contains all the information of its dynamics. To predict how India will do in the coming months, we predicted using each of these curves for the next 30 and 50 days (fig 6).

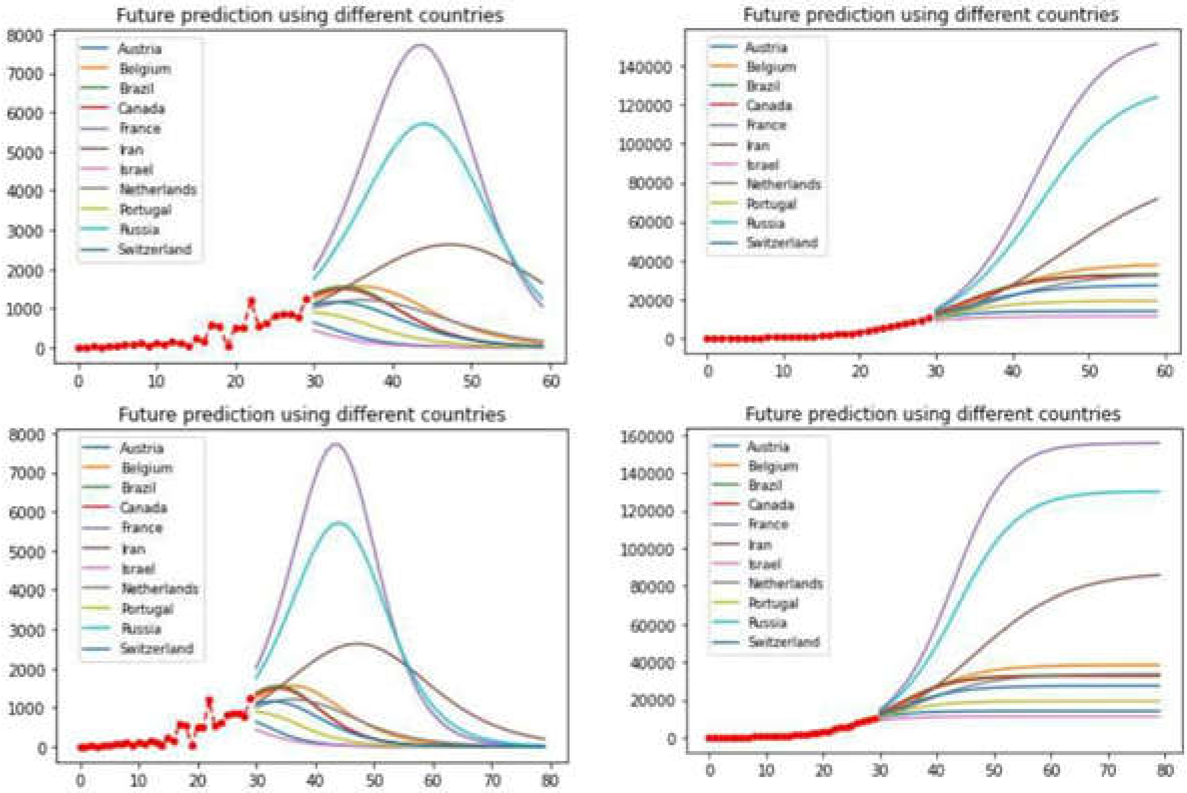


**Figure 6:** Represents the prediction for next 30 and 50 days, continuing the trend of specified countries for India (y-axis) and number of days (x-axis)

These are obtained by using the previous predicted value as an input for the next prediction and since we have the data for the countries till now, saturation is according to the current enforced steps in that country.

Each curve has its own saturation value, which gives us an idea where **we will end up if we follow that country's steps from this day onwards**. Since the model of each country is well fitted till the current cases in that country, this method gives us a better prediction till that value, which largely increases our capacity to predict, as only using the country's data, prediction for further days is not accurate due to reason reported above.

### Our Hypothesis

One other way of understanding this is that the prediction depends on two things, previous days and demographic/

administrative factors. Though the deep learning models did an excellent job in understanding the first relation, we hypothesize that all the information about the second is represented by [a, b, c] parameters that we have. Thus, to get the prediction for India, we learn it through the demographic/ administrative factors of other countries.

Here we will look at two important countries that form the **extreme cases namely Russia** (fig 7) **and Israel** (fig 8) **to introduce how different countries can be compared** leaving out the demographic features for now.

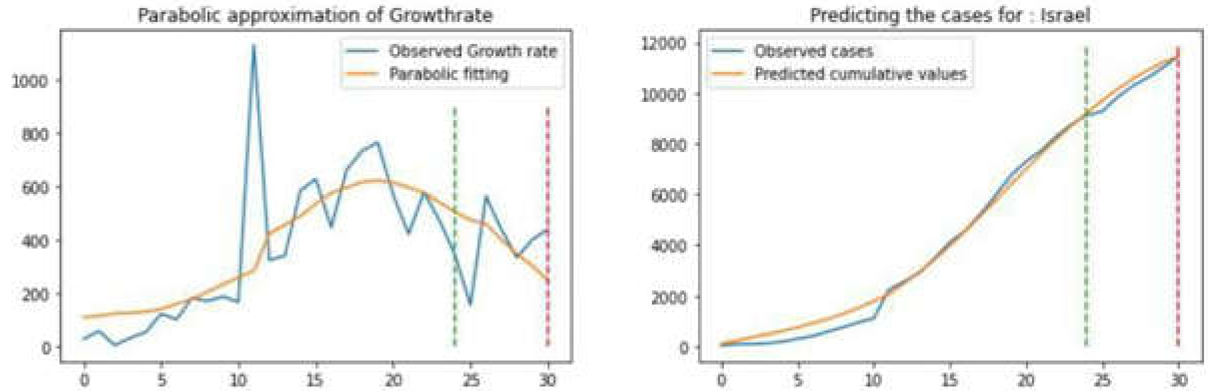


**Figure 7:** Represents prediction for Israel (y-axis) with number of days (x-axis), with analysis for the dynamics in which India is right now (greed-red dashed line)

Since Israel lies on the decreasing arm of parabola in the current scenario and the cases start to saturate after 10,000, this model gave us the lowest saturation limit. It will not be possible for us to trace this path as we are currently in a very steep rising phase. In depth analysis of what Israel did in recent days would give us really good insights.

For the case of Russia, it can be clearly seen that there is a steep rise in the domain in which we are right now, and thus this model gave us the steepest rise in another 30 or 50 days.

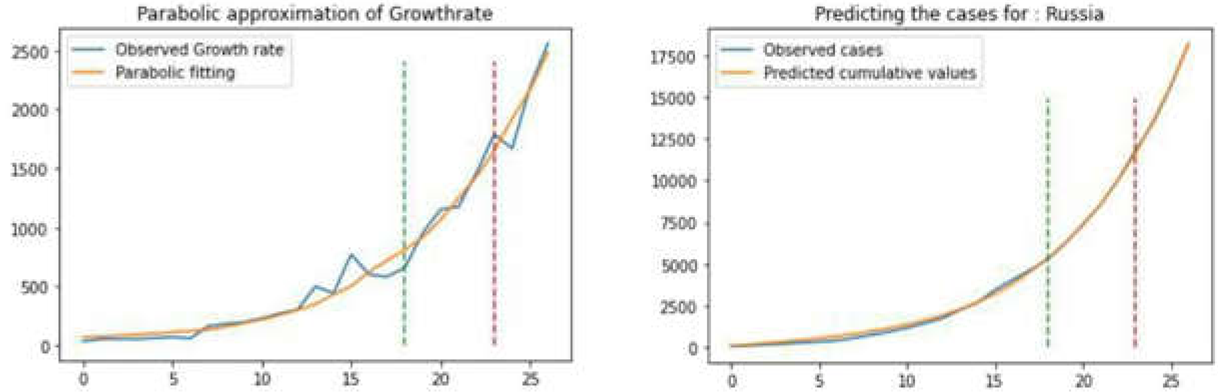


**Figure 8:** Represents prediction for Russia (y-axis) with number of days (x-axis), with analysis for the dynamics in which India is right now (greed-red dashed line).

This model and way of statistics gave us a way of knowing how our country can fare given what is the scenario now. We still lack the resources to compare the different actions taken by the countries so that we can suggest the optimum way we should choose to move ahead but we think that this way of comparing will definitely help the concerned authorities and if given resources, we have a proper plan of converting this into a better model, which is described in the future work.

Now, to show how this is useful, we will see how our prediction is, of India using these countries till 14th April, compared to the actual cases.

## 3. Further Analysis

We put this to test on the 11-day period, and thus we used the models we trained till 13th April data to predict the country's trend for next 11 days (fig 9.1). When we compared this response to the actual dynamics of cases in this period, we found that there are 6 countries whose prediction lies within the range of R2_score 0.8 (fig 9.2), thus using the models trained on these countries till 13th March data, we would have almost accurately predicted our number of cases. This again gives us a list of countries that we can look to, to predict how our country can evolve and thus we present the further 10, 30 days (fig 10) prediction using those countries.

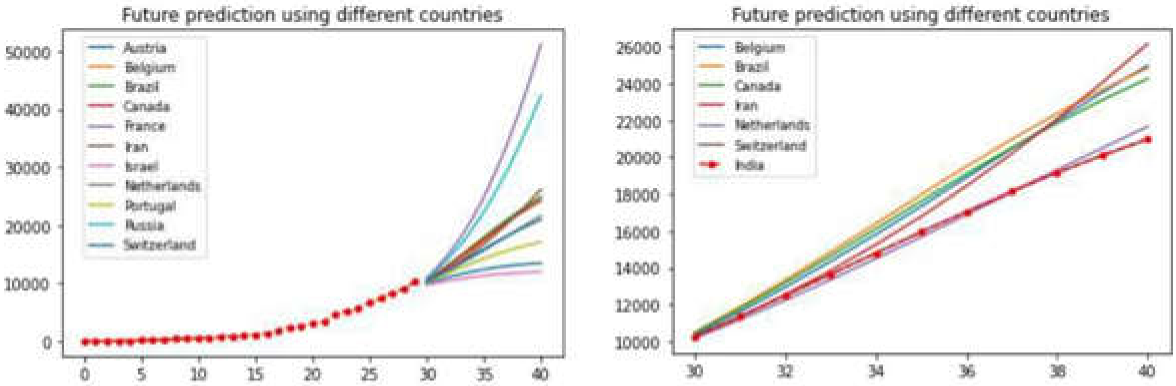


**Figure 9:** Represents 10 days future prediction for India's number of cases using different countries, and prediction of those countries whose prediction lies in the R2>0.8 of actual cases.

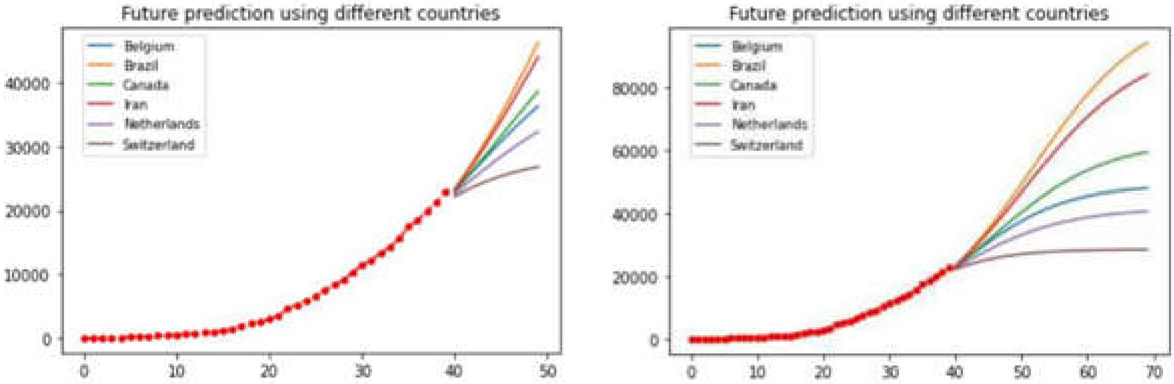


**Figure 10:** Represents number of cases that India could see according to different countries dynamics in further 10 and 30 days from 24th April

Now we can compare our growth (fig 11) in days between 23rd April-28th April.

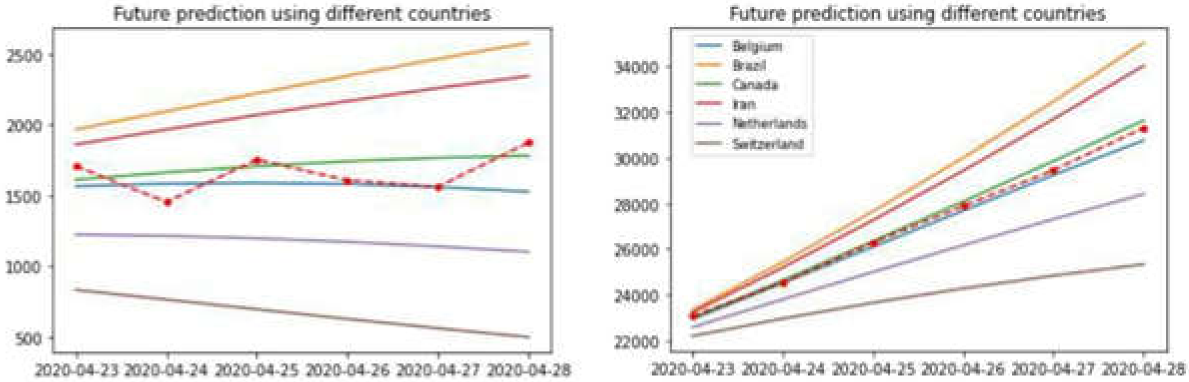


**Figure 11:** Represents India's growth rate and cumulative cases for the past 6 days predicted vs actual using six different countries trained using data till 23rd April

Results lie exactly within what we predicted and thus **Belgium and Canada** form the two strongest candidates to do the competitive study. This method clearly has an edge in long term forecasting, with results in excellent neighbourhood of prediction. This method of doing comparative study in two stages over last 15 days can be reiterated over evenly spaced intervals and we could get a fair estimate of how our country will evolve in upcoming days (fig 12) with least computation.

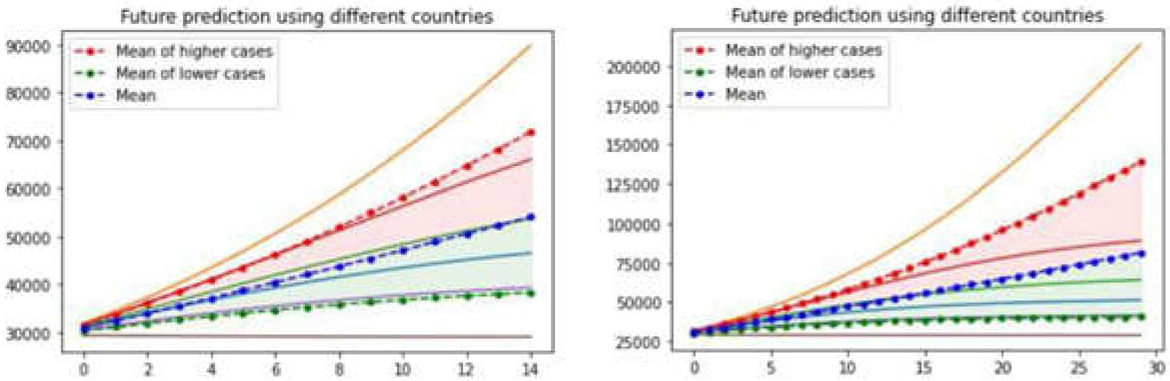


**Figure 12:** Represents prediction over 15 and 30 days using different countries along with mean prediction, mean prediction of higher-case countries and lower-case countries

## 4. Discussion

We suggest a novel approach of approximating the trend using other countries' data, which have had the same dynamics during their course of evolution. We conclude with suggesting the countries we are following, and countries we should follow. This can really help in modeling how this will turn up in further 30-50 days. Fitting a parabola over number of cases in different countries and then using those parameters to predict the number of cases in an evolving country follows the basis and shows excellent coherence with the results.

We further add to it an idea we are working on which can incorporate these demographic features as well of different countries but lack the proper resources to do so.

## 5. Future Work

We identified 10 different features that can be used to model why the country is following certain dynamics, i.e. [a, b, c] of the parabola can be shown as a function of these parameters. This list is neither exhaustive nor exclusive, and is just a stepping stone in what we are thinking can be used to build a model, taking these features of any country as the input and giving [a, b, c] of its growth as output. Around 30 countries that are in R2 > 0.8 for India can be used as training input for such a model, and there are parabola parameters as output. Then for different demographic conditions in our country we can get different parameters and thus different corresponding predictions that can take in sudden changes, government policies, population and temperature too into account.

We have a preliminary idea of the same ready but need further research to get parameters, check their correlation and converge the model to proper outputs. List of thought about parameters till now is as follows:

1) Population Density
2) Median age
3) Mean Temperature in March-April
4) Rate of tests being conducted
5) Number of hospitals and similar infrastructure per million
6) Health workers per million
7) Ratio of health workers affected by the virus
8) Days since Lockdown
9) Ratio of population in lockdown
10) Number of foreign nationals in Country in March

## References

[1] Chae, S.; Kwon, S.; Lee, D. Predicting Infectious Disease Using Deep Learning and Big Data. Int. J. Environ. Res. Public Health **2018**, 15, 159

[2] Chiou-Jye Huang, Yung-Hsiang Chen, Yuxuan Ma, Ping-Huan Kuo, Multiple-Input Deep Convolutional Neural Network Model for COVID-19 Forecasting in China, https://doi.org/10.1101/2020.03.23.20041608

[3] Joceline Lega, Heidi E. Brown, Data-driven outbreak forecasting with a simple nonlinear growth model, Epidemics, Volume 17, 2016, Pages 19-26, ISSN 1755-4365, https://doi.org/10.1016/j.epidem.2016.10.002.

[4] https://github.com/CSSEGISandData/COVID-19

[5] https://github.com/datasets/covid-19

[6] https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467 b48e9ecf6

[7] https://docs.google.com/spreadsheets/d/e/2PACX-1vSz8Qs1gE_IYpzlkFkCXGcL_BqR8hZieWVi-rphN1gfrO3H4lDtVZs4kd0C3P8Y9lhsT1rhoB-Q_cP4/pubhtml#

[8] https://www.covid19india.org/

[9] COVID-19 and Italy: what next? Andrea Remuzzi, Giuseppe Remuzzi

[10] https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model