A TRANSPARENT, OPEN-SOURCED SIRD MODEL FOR COVID19 DEATH PROJECTIONS IN INDIA

Ananye Agarwal
Computer Science and Engineering
Indian Institute of Technology, Delhi
cs1170326@iitd.ac.in
ananayagarwal@gmail.com

Utkarsh Tyagi
Electrical Engineering
Indian Institute of Technology, Delhi
ee3170550@iitd.ac.in
utkarshtyagi99@gmail.com

May 30, 2020

ABSTRACT

As India emerges from the lockdown with ever higher COVID19 case counts and a mounting death toll, reliable projections of case numbers and deaths counts are critical in informing policy decisions. We examine various existing models and their shortcomings. Given the amount of uncertainty surrounding the disease we choose a simple SIRD model with minimal assumptions enabling us to make robust predictions. We employ publicly available mobility data from Google to estimate social distancing covariates which influence how fast the disease spreads. We further present a novel method for estimating the uncertainty in our predictions based on first principles. To demonstrate, we fit our model to three regions (Spain, Italy, NYC) where the peak has passed and obtain predictions for the Indian states of Delhi and Maharashtra where the peak is desperately awaited.

1 Introduction

At the higher level, the model implements the SIRD Compartmental Model in Epidemiology with inputs from the actual social distancing behaviour of the people in the country, which allows us to vary the Reproduction Rate (R_0) for the infection by capturing details of the policies being implemented with respect to lockdowns in the country.

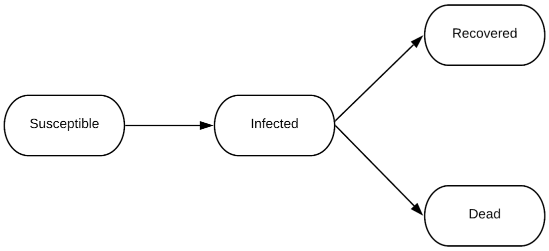

Figure 1: Block Diagram for SIRD Model

2 Related Work

We list down two models being referred to in the United States and their challenges, and then we discuss how we overcome them.

IHME Model After getting cited by the United States Government for the COVID19 projections, the IHME model [1] has received a lot of attention from the public and professional community. It relies on data from China and Italy for training the model parameters and extend the model to the United States. One of the criticisms for the model is that it uses statistical definitions to make predictions. It cannot give information on the actuals about the disease spread due to departure from epidemiological theory. It has consistently under-performed compared to the actual statistics, and needs to be re-updated after every few weeks.

Multi-Compartment Model Overcoming the limitations of the statistical models like IHME Projections, Stanford published a nine compartmental model[2], which explicitly tracks nine compartments, including exposed, asymptomatic, pre-symptomatic, symptomatic, hospitalized, and recovered. Similar model was developed in India as INDSCI-SIM model. [3] The criticism to the extensive epidemiological model is that they have a large number of tunable and learnable parameters. Data handling for such models becomes unreliable as amount of input data required increases. Since the data provided by the government becomes unreliable as the cases increase, the modelling in such cases becomes unreliable.

PRACRITI Model by M3RG, IIT Delhi The model by M3G Group of IIT Delhi[4] is an extension of the SIER Model that creates Adaptive, Interacting, Cluster-based, Susceptible, Exposed, Infected, Removed compartments for district level populations. This model suffers from limitations as stated for both previous models. They have added a lot of mathematical parameters with do not have physical meaning, and thus cannot be checked against real-world data. With additional migration term, they assume the migration statistics irrespective of the movement of the migrants and their social distancing measures. Next, the model doesn't perform well as on simple inspection, the R_0 for Mumbai is reported as 0.73, while same value for the State of Maharashtra is 1.33.

3 Novelty

Our model differs from the models discussed above in at least three key ways, which collectively result in different forecasting behavior.

Modelling on Daily Deceased Data: We use daily deceased data for time-series forecasting of the COVID19 infections, and model the number of infected people based on that data. The benefits of single time-series over models also using Infected time series data is that their mortality rate depends on the scale of testing and reporting by the relevant government agencies. We bypass the uncertainty of testing levels by using death data because it is more reliable even in cases where testing is low.

Limited Uncertainty: The model implements a four compartment SIRD based model where we vary reproduction number (R_0) with respect to the social distancing measures. This gives us flexibility to monitor real-time policy changes in the data. This also makes the model region agnostic and can be implemented on national and state levels, all over the world.

Transparency: Since our model is based on actual data and relates to the epidemiological theory, we can measure daily reproduction number, which allows us to do a sanity check with respect to the current situation. This can be used to compare policy-level effects in mitigating the spread of the virus by comparing the variation in the reproduction number.

4 Description of the Model

4.1 Mathematical formulation

An SIRD model is described by the following coupled differential equations,

$$\frac{dS}{dt} = -\frac{RIS}{T_{inf}} \tag{1}$$

$$\frac{dI}{dt} = \frac{RIS}{T_{inf}} - \frac{I}{T_{inf}} \tag{2}$$

$$\frac{dX}{dt} = \gamma_X I \tag{3}$$

Here, S, I, X are respectively the fraction of the population that is susceptible, infected and deceased. We do not consider the number of people who have recovered because we choose not to estimate case fatalities due to the unreliable number of case counts.

T_{inf} denotes the median amount of time a person stays infectious, γ_X is average number of people who die from COVID19 in a day as a fraction of the total number of active cases on that day. R is the reproduction number of the disease which measures the average number of people an infected person transmits the disease to. $R > 1$ implies that the case count rises over time while $R < 1$ implies that the case count diminishes over time with the rate of spread being determined by R. Note that in general R varies with time depending on the extent of social distancing practiced.

Note that γ_X and T_{inf} are time-invariant properties of the disease. Therefore, we can bound these parameters to within a small interval by looking at how COVID19 has progressed in other places. We then see that all the information about the progression of the disease lies in a single parameter - R. This is the major advantage of using a simple model. We do not have to deal with lots of uncertain parameters that influence the final curve in unpredictable ways, we can focus on estimating R as best we can. Further, R is a well-established measure of disease spread in epidemiological literature which means there are already many existing estimates of R which our model can leverage.

4.2 The parameter R

R is a function of time and depends on (among other things), the lockdown and social distancing measures adopted by each country. To estimate R we leverage open-source, real-time social distancing data published by Google [5], which allows us to model various mitigation measures by just two parameters as described below. While the social mobility data does not account directly for various measures such as contact tracing and mask usage, we nonetheless postulate that the timing of these measures is correlated with the timing of social distancing measures indicated by the mobility data.

The data, available in aggregated form, shows how the number of visitors who go to (or spend time in) categorized places change compared to pre-COVID days. A baseline day represents a normal value for that day of the week. The baseline day is the median value from the 5-week period Jan 3 – Feb 6, 2020. The places are categorized into Retail and Creation, Grocery and Pharmacy, Parks, Transit Stations, Workplaces, and Residential.

Additionally, for a sanity check, we looked at the smartphone penetration in the country to validate the model. The report by StatCounter [4] suggests that Android based smartphones constitute more than 95% of all smartphones being used in the country in April 2020. With this size of market share, the open source data-model performs well.

We construct a social distancing covariate $s(t)$ from the changes in social mobility at various locations

$$s = \Delta I_r(t) + \Delta I_g(t) + \Delta I_p(t) + \Delta I_t(t) + \Delta I_w(t)$$

where each of the terms on the RHS above denote percentage change from baseline in mobility at the following locations

- $\Delta I_r(t)$ - retail and recreation
- $\Delta I_g(t)$ - grocery and pharmacy
- $\Delta I_p(t)$ - parks
- $\Delta I_t(t)$- transit stations
- $\Delta I_w(t)$ - workplaces

Note that we ignore residential mobility data as residential mobility does not contribute in a spread of the disease. Next, we smooth the covariate $s(t)$ by applying a Savitzky–Golay filter followed by convolution with a localized Gaussian multiple times. The goal is to smooth out weekly variations in the data but not distort the overall profile of the curve. This gives us $\mathcal{S}(t)$, the smooth social distancing covariate.

Since we only care about the timing of social distancing measures, to relate $\mathcal{S}(t)$ to R we introduce two parameters R_{min}, R_{max}, the R values when $\mathcal{S}(t)$ is minimum and maximum respectively. We then define R as a linear interpolation function of $\mathcal{S}$ between these two values. Mathematically,

$$\frac{R(t) - R_{min}}{R_{max} - R_{min}} = \frac{\mathcal{S}(t - \delta_{sd}) - \mathcal{S}_{min}}{\mathcal{S}_{max} - \mathcal{S}_{min}}$$

where $\mathcal{S}_{max}, \mathcal{S}_{min}$ are the global maximum and minimum values of $\mathcal{S}(t)$. Further, we introduce a fixed lag δ_{sd} which equals the median time from infection to death. This is because $\mathcal{S}(t)$ influences the number of infections at time t, the

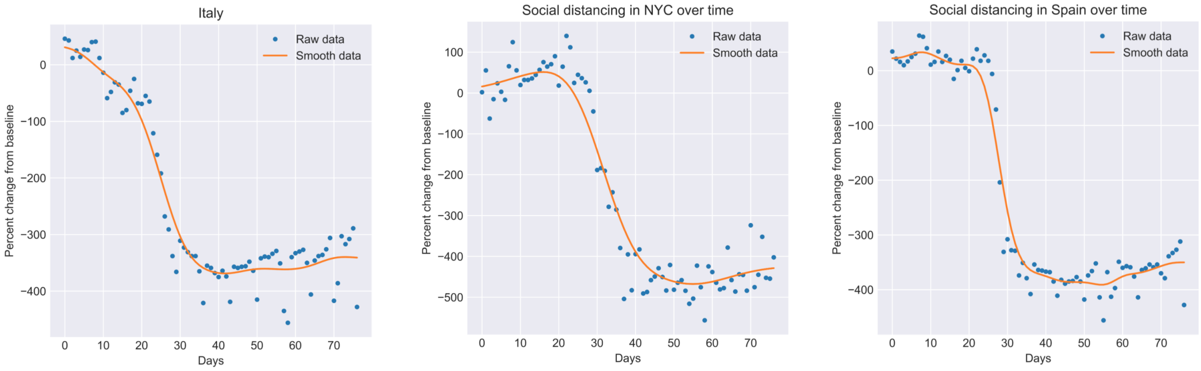

Table 1: Raw and smooth social distancing data for three different regions from 15 Feb

effect on deaths is seen only later. Where social distancing data is not available, we naively extrapolate the existing data into the future as well as the past. Concretely, we assume that past values equal the earliest value we know and future values equal the latest value. This amounts to assuming that existing social distancing measures will continue into the future. This assumption can be altered as we learn more about the disease and mitigation strategies in the future.

4.2.1 Fitting the model

We solve the differential equations using a simple iterative procedure where the values of the next day are determined by the values of the previous day.

$$S_{t+1} = S_t - \frac{R_t I S_t}{T_{inf}} \tag{4}$$

$$I_{t+1} = I_t + \frac{R_t I S_t}{T_{inf}} - \frac{I_t}{T_{inf}} \tag{5}$$

$$X_{t+1} = \gamma_X I_t \tag{6}$$

We avoid more complicated numerical techniques like the Runge-Kutta methods because they proved to be too computationally intensive to fit a large number of models. Additionally, for our purposes the above recurrences yield a reasonable approximation.

Note that to solve the system of differential equations above we need to specify an initial condition. In particular we need to specify initial values for time t, and each of S, I, X. Since the set of differential equations $(1) - (3)$ is valid at all points of time we can arbitrarily choose a starting point.

We start the model just before we get the first death. Obviously, $S_{t_0} = 1$, $X_{t_0} = 0$. We choose $I_{t_0} = \frac{1}{\gamma_X P}$ where P is the population. This choice implies that at day 1, there will be exactly one death. Since real death counts are discrete, we choose t_0 in a narrow interval around where the actual death count start to rise.

To prepare the daily death counts, we obtain raw death counts from two sources - John Hopkins [6] and the covid19india.org [7], a volunteer-driven tracker project. We then smooth this death count using a combination of Savitzky–Golay and Gaussian convolution filters. Care needs to be taken to not distort the peak too much as with a large amount of smoothing the peak tends to decrease in height.

Finally, to fit the data we do a fine-grained brute force grid search over the possible parameter values we provide and obtain a prediction with the lowest mean squared loss. In general, we fix $\gamma_X = 1.6\text{e}{-3}$ (this implies a mortality rate of 0.8%), $\delta_{sd} = 23$, vary t_0 within a small margin near the beginning of the death count curve and vary R_{max} from $1.4 - 2.8$ and R_{min} from $0.7 - 0.95$.

Predicting the peak (height and position) are quite tricky because the beginning of the curve looks quite similar for different values of R. Further, the peak can be quite sensitive to the values of R_{max} and R_{min}. R_{min} is especially hard to estimate because it depends on the death count close to or after the peak, after social distancing measures have been put into place. We discuss these issues in greater detail in the following section.

4.3 Uncertainty Analysis

Let $\mathbf{M}$ be the random vector corresponding to the choices for (variable) parameters in the model. Then, the density function $f(\mathbf{m})$ constitutes a prior on the choice of these parameters. As an approximation, we assume uniform priors on the parameters (this can in principle be extended to other priors). This is reasonable because from our knowledge of other countries, we can place bound R quite confidently whereas pinpointing a single value of R is very hard.

Further, we assume that the data $y_t \sim h(t; \mathbf{m}) + \mathcal{N}(0, \sigma^2)$, where h is the hypothesis (SIRD model), y_t is the observed deaths on day t and $\mathbf{y}$ is the vector of observed deaths. Note that

$$\mathbb{P}(\mathbf{M} = \mathbf{m} \mid \mathbf{y}) = \mathbb{P}(\mathbf{y} \mid \mathbf{M} = \mathbf{m})\mathbb{P}(\mathbf{M} = \mathbf{m})$$

We can assert that we only include parameters in our confidence interval which have probability atleast ϵ

$$\mathbb{P}(\mathbf{M} = \mathbf{m} \mid \mathbf{y}) \geq \epsilon \implies \mathbb{P}(\mathbf{y} \mid \mathbf{M} = \mathbf{m}) \geq \frac{\epsilon}{\mathbb{P}(\mathbf{M} = \mathbf{m})} \tag{7}$$

$$\log\left(\mathbb{P}(\mathbf{y} \mid \mathbf{M} = \mathbf{m})\right) = \sum_{t=0}^{t_{max}} \left[\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(y_t - h(t; \mathbf{m}))^2}{2\sigma^2}\right] = -(t_{max} + 1)\left[\log\left(\sqrt{2\pi}\sigma\right) + \frac{L(\mathbf{y}, \mathbf{m})}{2\sigma^2}\right]$$

where $L(\mathbf{y}, \mathbf{m})$ is the average root mean squared loss between $\mathbf{y}, \mathbf{m}$. This means that (7) is equivalent to

$$-(t_{max} + 1)\left[\log\left(\sqrt{2\pi}\sigma\right) + \frac{L(\mathbf{y}, \mathbf{m})}{2\sigma^2}\right] \geq \log\frac{\epsilon}{\mathbb{P}(\mathbf{M} = \mathbf{m})}$$

Simplifying this we get,

$$L(\mathbf{y}, \mathbf{m}) \leq \frac{2\sigma^2}{t_{max} + 1}\log\frac{\mathbb{P}(\mathbf{M} = \mathbf{m})}{\epsilon} + 2\sigma^2\frac{1}{\sqrt{2\pi}\sigma}$$

Note that because of the uniform prior the RHS is independent of $\mathbf{m}$. Further, for a theoretical perfect fit, $\epsilon = 1$ and $\mathbb{P}(\mathbf{M} = \mathbf{m}) = 1$, making the first term zero. Therefore, we can interpret the second term $2\sigma^2\frac{1}{\sqrt{2\pi}\sigma}$ as the minimum loss (or the loss of the best fit curve). This gives us a metric for choosing admissible values of the parameter $\mathbf{m}$.

$$L(\mathbf{y}, \mathbf{m}) \leq \frac{\alpha}{t_{max} + 1} + L(\mathbf{y}, \mathbf{m}^*)$$

where α is a constant we choose and $\mathbf{m}^*$ are the best fit parameters. Having obtained a set of values of $\mathbf{m}$, we obtain the corresponding curves for them and plot the minimum and maximum predictions for all these curves to obtain a confidence interval. Note that the acceptable range of $L(\mathbf{y}, \mathbf{m})$ grows smaller as t_{max} increases i.e. as we get more data. This agrees with our intuition, which says that as we get more data the confidence interval should become narrower (for fixed t, α)

In practice, we start with a conservatively high value of α (=200). As we get more data, we increase α if the actual values fall outside the confidence interval.

5 Results

Here we present results for three different regions whose death count peaks have passed. All three - Spain, NYC, Italy were badly effected by coronavirus as India is likely to be. We use these curves to validate the values we have chosen for the fixed parameters.

Note that more detailed curves are present in the Github repository. The curves are self-explanatory. It is worth noting that as we expect, with more data the uncertainty interval narrows and converges to the observed data.

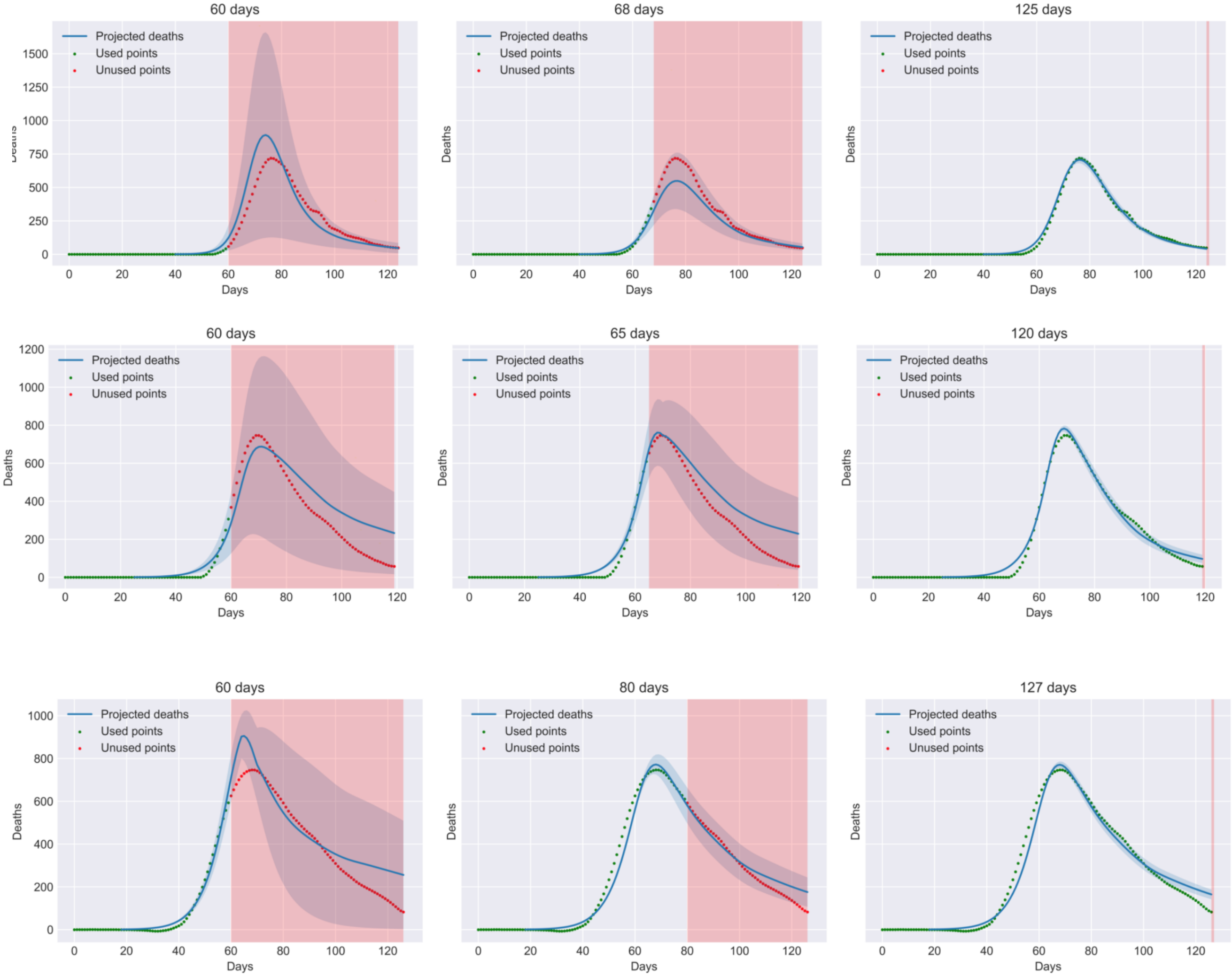

Table 2: Predictions with uncertainty intervals for three different regions New York City (NYC), Spain, and Italy (from top to bottom). In each figure, the red area contains points the model has not been fitted on and the shaded blue region is a confidence interval.

Breakpoint	**Train Loss**	**Population**	_γ_X	_$\mathbf{I}_{init}$	**Offset**	_$\mathbf{R}_{max}$	_$\mathbf{R}_{min}$
60	6.0678	8.4 mil	1.6e-3	7.4e-5	40	2.0571	0.80
68	15.9692	8.4 mil	1.6e-3	7.4e-5	40	2.3429	0.95
125	18.4782	8.4 mil	1.6e-3	7.4e-5	40	2.4000	0.8833

Table 3: Model Parameters for NYC

Breakpoint	**Train Loss**	**Population**	_γ_X	_$\mathbf{I}_{init}$	**Offset**	_$\mathbf{R}_{max}$	_$\mathbf{R}_{min}$
60	26.0657	46.9 mil	1.6e-3	1.3e-5	25	1.9474	0.95
65	28.4373	46.9 mil	1.6e-3	1.3e-5	25	1.9842	0.87
120	28.5387	46.9 mil	1.6e-3	1.3e-5	25	1.9842	0.84

Table 4: Model Parameters for Spain

5.1 Predictions

We now include predictions for two critical regions in India - Delhi and Maharashtra, both badly affected by the virus.

Breakpoint	**Train Loss**	**Population**	_γ_X	_$\mathbf{I}_{init}$	**Offset**	_$\mathbf{R}_{max}$	_$\mathbf{R}_{min}$
60	22.9539	60.4 mil	1.6e-3	1.1e-5	18	1.9895	0.5
80	41.9706	60.4 mil	1.6e-3	1.1e-5	18	1.9158	0.8868
127	40.9382	60.4 mil	1.6e-3	1.1e-5	18	1.9158	0.8611

Table 5: Model Parameters for Italy

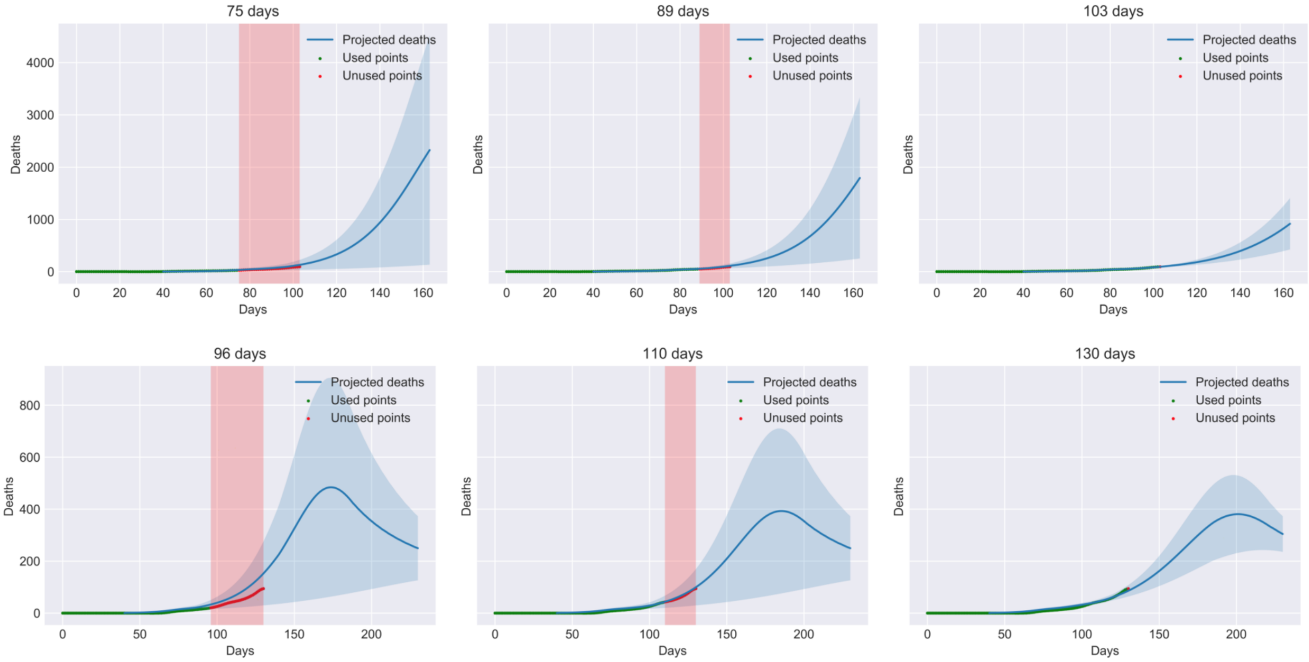

Table 6: Predictions with uncertainty intervals for Maharashtra and Delhi (from top to bottom). In each figure, the red area contains points the model has not been fitted on and the shaded blue region is a confidence interval. We give projections for the death counts into the future in both cases.

Breakpoint	**Train Loss**	**Population**	_γ_X	_$\mathbf{I}_{init}$	**Offset**	_$\mathbf{R}_{max}$	_$\mathbf{R}_{min}$
80	4.1892	114.2 mil	1.6e-3	5e-6	40	1.6316	1.1000
89	3.8369	114.2 mil	1.6e-3	5e-6	40	1.6316	1.1105
103	4.0742	114.2 mil	1.6e-3	5e-6	40	1.5737	1.2158

Table 7: Model Parameters for Maharashtra

Breakpoint	**Train Loss**	**Population**	_γ_X	_$\mathbf{I}_{init}$	**Offset**	_$\mathbf{R}_{max}$	_$\mathbf{R}_{min}$
96	2.5622	19 mil	1.6e-3	3.3e-5	40	1.4	1.1316
110	3.0304	19 mil	1.6e-3	3.3e-5	40	1.4	1.1842
130	3.4631	19 mil	1.6e-3	3.3e-5	40	1.4	1.1947

Table 8: Model Parameters for Delhi

Note that the uncertainty intervals for Maharashtra in the beginning of the curve are very high. This indicates that the model does not fit well to the initial part of the curve. This might be a consequence of (a) the fact that Maharashtra is a large state and different parts of the state are affected differently by the virus (a model fit to Mumbai would perform better) (b) The timing of the drop in $\mathcal{S}(t)$ does not track well the timing of prevention measures taken in the state.

On the other hand, the model performs quite well on Delhi. This is likely because Delhi is much more homogeneous in terms of demographics and is more well-connected. This means that Google mobility data is likely to reflect well how much people are social distancing. Note that the current projections in Delhi assume that social distancing will continue at lock-down levels. This is likely to not be the case as Delhi has started re-opening. Nevertheless, the government is still attempting to aggressively identify and quarantine so-called containment zones.

It is worth noting that a clear conclusion from the data is that even during the lock-down which has been called one of the strictest in the world, R_{min} remained above 1, unlike in other countries. This can indeed be seen from our model as well which predicts $R_{min} < 1$ for Italy, NYC and Spain but $R_{min} > 1$ for Maharashtra and Delhi. Now that the country is re-opening R can only be expected to increase further. It is hoped that through this work we can emphasize the urgency with which the India needs to find an effective strategy to contain the virus.

5.2 Further work

Based on our discussion in the previous sections, we can se the following directions in which the model can be improved

- **Improving quality of R estimation** With the wide usage of the Aarogya Setu app, the government has accurate raw data available for people's movement patterns. If we can acquire this data through official channels, we can further improve our R estimates.
- **Constructing an online dashboard** We can construct an online dashboard which shows projections with uncertainty intervals in real-time for all districts in India. Further, we can allow the user to transparently adjust R to understand how critical social distancing is to contain the spread.
- **Improving uncertainty estimation** Currently we choose α based on empirical conditions. Which is to say, we run the model on many different countries and choose α for which a large majority of predictions fall within the confidence interval. Can we choose α in a more principled manner?

References

[1] IHME | COVID-19 Projections.

[2] Potential Long-Term Intervention Strategies for COVID-19.

[3] A state-level epidemiological model for India: INDSCI-SIM -. Library Catalog: ZoteroBib.

[4] R. Ravinder, Sourabh Singh, Suresh Bishnoi, Amreen Jan, Abhinav Sinha, Amit Sharma, Hariprasad Kodamana, and N. M. Anoop Krishnan. An Adaptive, Interacting, Cluster-Based Model Accurately Predicts the Transmission Dynamics of COVID-19. preprint, Epidemiology, April 2020.

[5] Google LLC . Google COVID-19 Community Mobility Reports.

[6] CSSEGISandData. CSSEGISandData/COVID-19, May 2020. original-date: 2020-02-04T22:03:53Z.

[7] covid19india/api, May 2020. original-date: 2020-03-21T05:05:50Z.