Report on Covid19 India data,
ISRC Mathematical Modelling Group

Sarika Bhattacharyya, Pinaki Chaudhuri, Sujata Ghosh, Abhishek Mandal,
Ujjwal Kumar Nandi, Bhishmaraj Selvamani, Shiladitya Sengupta, R. Shankar and Prateek Sharma

April 20, 2020

1 Introduction

The overall aim of this sub-group is to study and analyse the data from tests. While this data is insufficient to answer many pressing questions, this effort is motivated by an ancient Indian saying, "If you play enough with the data, it will start talking to you". The aim is to understand the available data and see what useful information can be inferred from it. We are mainly concentrating on the data from India, however whenever a comparison with international data is useful, we analyse those too.

1.1 General remarks

We are analysing the data coming from the tests. This has the following obvious limitations:

1. The number of tests is limited. On 8^{th} April, around 1.3 lakh tests have been conducted (source ICMR). This is around 0.01% of the population. However, since the infections, as of now, are confined to local regions, the sampling ratio will be much better than 0.01%. How much better ? The answer is not clear yet.

2. The tests are not random but are biased towards the more serious cases and towards established hotspots. So the data is representative of the population satisfying the (changing) test criteria. How much is this population and how is it distributed ? The answer is not clear yet.

Basically, since, as yet, the infections are far from uniformly distributed, the interpretation of the parameters we extract by fitting nation wide or state level data of confirmed cases to mathematical functions has to be done with care.

1.2 Data sources

We have used data from:

Data ID	Source URL
D1	https://ourworldindata.org/coronavirus
D2	https://github.com/CSSEGISandData/COVID-19
D3	https://www.covid19india.org
D4	https://covindia.netlify.com
D5	https://en.wikipedia.org/wiki/Timeline_of_the_2020_coronavirus_pandemic_in_India#April_7-8

1. D1 gets data from the European Center for Disease Prevention and Control (ECDC).

2. D2, CSSE expands to Center for Systems Science and Engineering at Johns Hopkins University.

3. D3, Covid19india.org (dataset 3) is run by "a group of dedicated volunteers who are curating the news coming from state bulletins, press releases and news houses. We extract the details, like patient relationship to other patient (to identify local and community transmissions), travel history and status. We never collect or expose any personally identifiable data regarding the patients."

4. D4, covindia.netlify.com is run by "a dedicated team of students and professors from Mahindra Ecole Centrale who are committed to keeping everyone as informed as we possibly can."

5. D5 reports on data is from the Ministry of Health and Family Welfare, Government of India (MoHfW).

Here, we note that similar to D3, it would be useful if D5 provides detailed data on patients and their contact networks.

1.3 Summary

Our main results are:

1. **National level analysis**
   (a) A comparison of different data sets in section 2.2 shows systematic inconsistencies between the MoHFW data and the rest.
   (b) An analysis of the time evolution of the growth rates of the total number of confirmed cases in India in section 2.3 shows effects of the travel ban and the lockdown.
   (c) The growth rate of confirmed cases has been decreasing from around the begining of April.
2. **State level analysis**
   (a) The growth rates of the confirmed cases in the five states analysed in section 3.1 show significant variation. Currently the doubling time varies from 4.5 days in Maharashtra to 11.3 days in Kerala.
   (b) The growth rates of all these five states are generally decreasing after around April 2.
   (c) An analysis of the ratio of the number of confirmed cases where contact tracing was not possible to those where it was possible done in section 3.2 shows that, on April 9, 8 of the 17 states that had reported cases have this ratio greater than 10. This indicates that community transmission had set in strongly in about half the states that have confirmed cases about a week ago.
   (d) In Maharashtra, the above ratio was about 125 on April 9, the highest among all states. This may be the primary reason that it also has the highest growth rate of confirmed cases.

The overall impact of the travel ban and lock down can be gauged from Figure 1

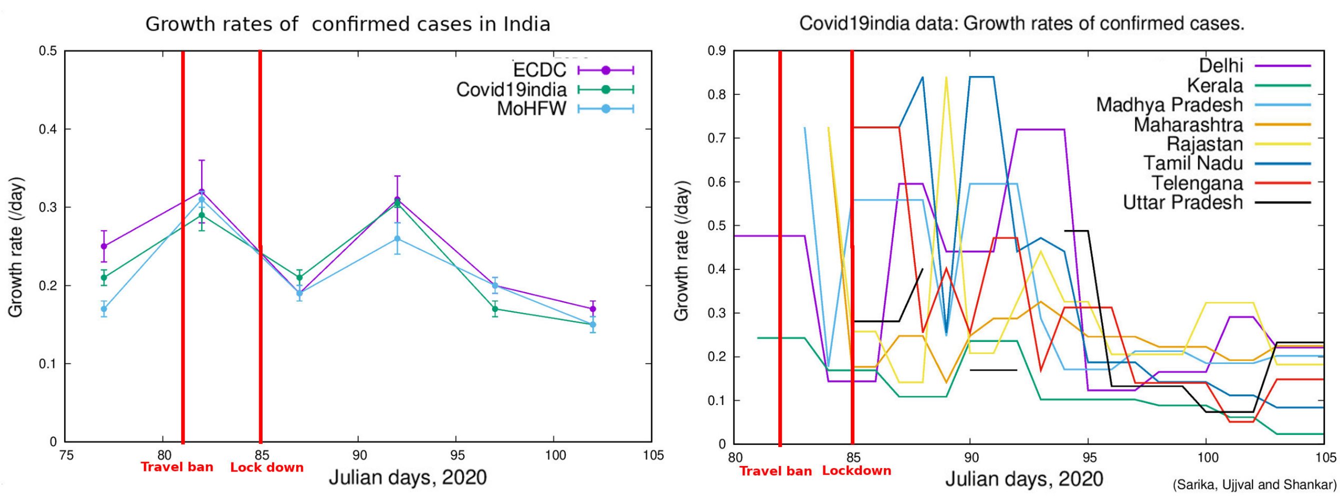

Figure 1: The growth rates of the confirmed cases in India estimated using three different data sets is shown in the left panel.The right panel shows growth rates of eight states

There is an initial dip in the national data and that of several states just after the travel ban was imposed. After that, till the begining of April, the growth rates of the individual states fluctuate and the national growth rate increases. After the begining of April the national growth rate and the overall trend of the state growth rates have benn decreasing. However, only Kerala and Tamil Nadu show a consistent decrease after April 5 (julian day 96). It seems natural to conclude that the decrease of growth rates in April is a consequence of the lock down, it seems to have taken about a week to have a significant effect. Of the states analysed only Kerala and Tamil Nadu seem to been able to use the first phase of the lock down to steadily bring down their growth rates. The other six seem to be still struggling.

We repeat the caveat that all the above statements are based on the growth rates of the confirmed cases which may or may not reflect the trends of the actual number of cases.

2 National level analysis

2.1 Comparison of India with some other countries

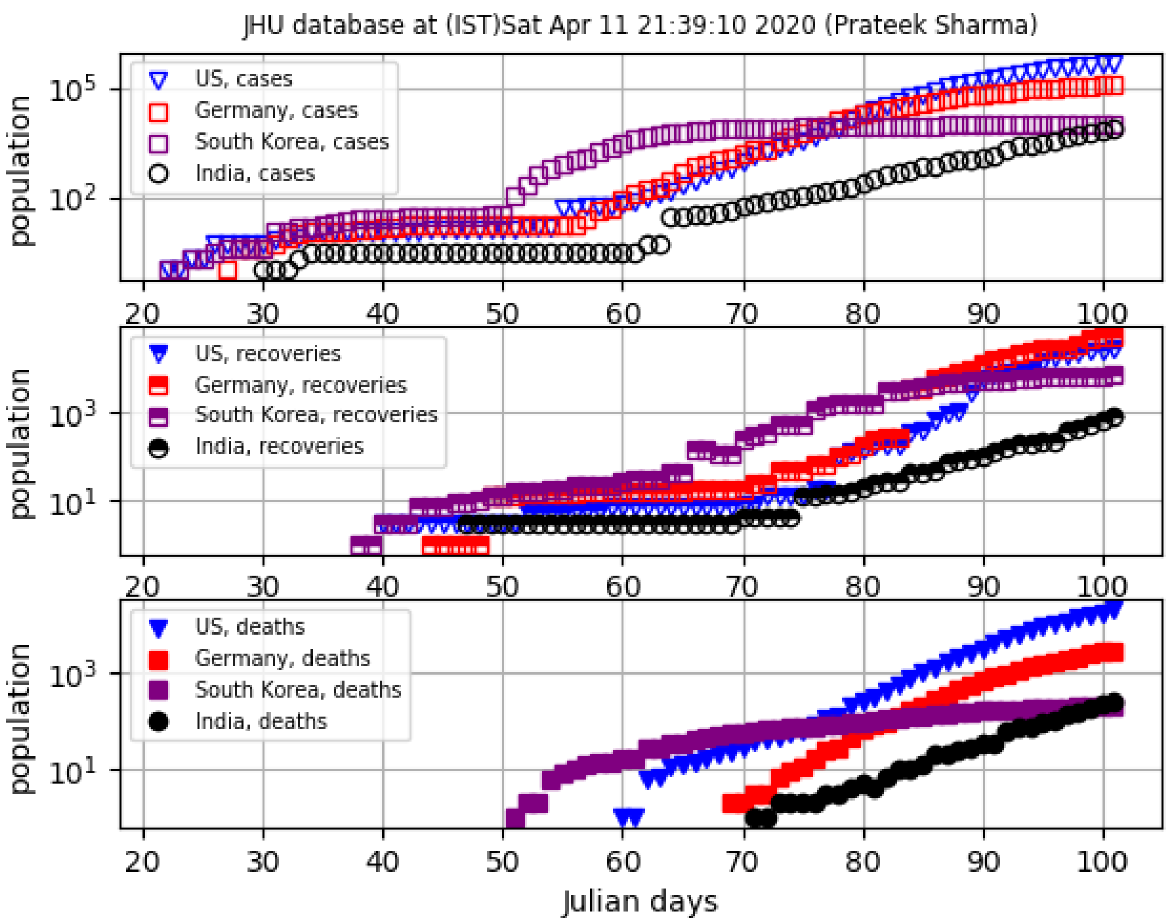

Figure 2: Number of cases, recoveries and deaths versus time from the JHU database for some countries.

Before going in detail over the data from India, it is useful to look at the aggregate numbers for India and compare them with some other countries. Note that South Korea and Germany have tested a much higher fraction of their population compared to US and India.

2.2 A comparison of different data sets

In this section we compare the time series of confirmed cases from datasets D1,D2,D3,D4 and D5. The time series of recovered and deceased cases from D2,D3 and D5.

2.2.1 Questions

One simple question is being asked. How much is the variation between these data sets.

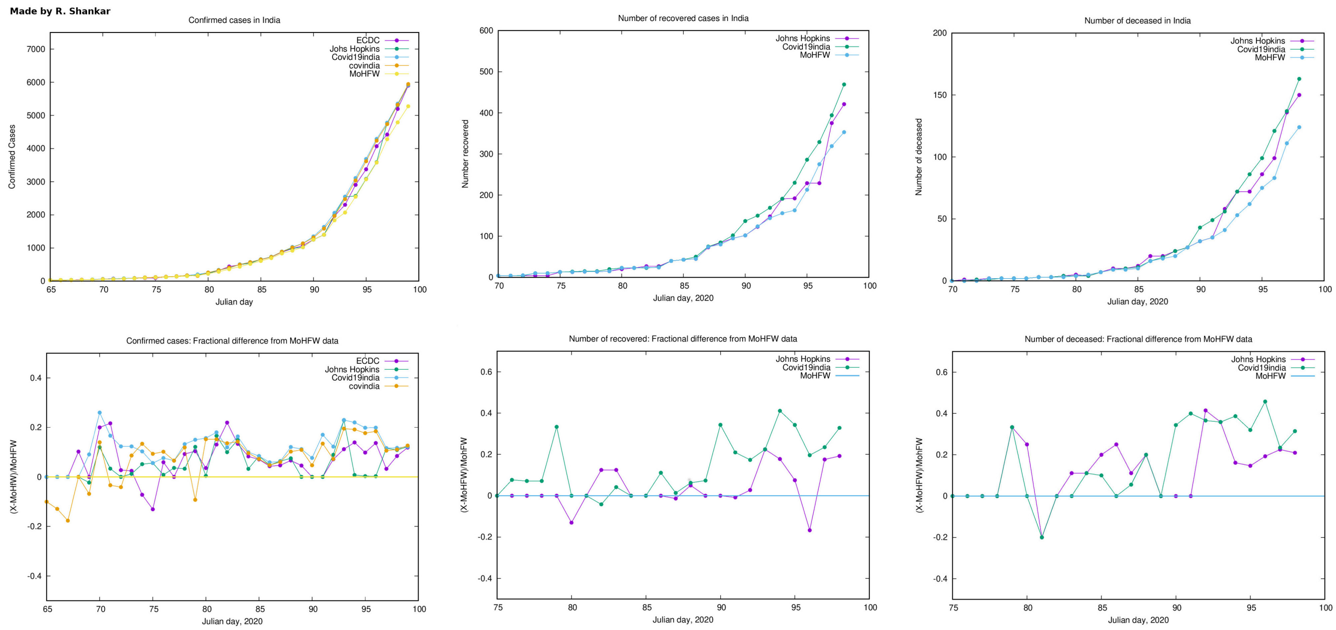

Figure 3: The numbers of confirmed, recovered and deceased cases for the different datasets are plotted in the upper panel. The fractional difference with respect to the MoHFW data, as defined in the main text are shown in the lower panel.

2.2.2 Methodology and results

To quantify the variability, the fractional difference, f, between the MoHFW data (D5) and the rest was computed and plotted. f is defined as

$$f(t) \equiv \frac{(N_n(t) - N_5(t))}{N_5(t)}$$

where N_n is the number given by D_n. Note, D_5 is the MohWF dataset.

The data for the confirmed, recovered and deceased cases for all the data sets are shown in the upper panel of Figure 3. The fractional difference, f, for the three cases is shown in the lower panel.

2.2.3 Interpretation and discussion

1. There is a fair amount of spread in the absolute numbers, especially in late March and April. However the numbers themselves are also growing.
2. The MoHWF numbers are more or less uniformly lower than the rest. This implies that is not a statistical error but a systematic one.
3. The fractional differences between the MoHWF data and the other sets for the confirmed cases fluctuates in time in a the range 0-20%.
4. The fractional difference between the MoHFW data and the other sets for the recovered and deceased cases fluctuates in the range 0-20% till around March 27 (JD 87), it increases to after that to as much as 40%. For the last week or so, the fractional difference for the Johns Hopkins set fluctuates about 20% and that of the Covin19india about almost 40%.
5. It is important to understand the reasons for the sytematic inconsistencies discussed above.

2.3 Growth rates and their changes

In this section we analyse the time series of the confirmed, recovered and deceased in terms of their growth rates or equivalently the doubling times. We would like to stress here that the parameters we extract are to be interpreted with care as they are averages over an extremely heterogeneous system.

2.3.1 Questions

What are the growth rates and changes in them ? Can we relate the changes to the major mitigatory measures, namely the travel ban and the lock down.

2.3.2 Methodology and results

If the number of cases is growing exponentially, the growth rate, κ, is defined by, $N(t) = N_0 e^{\kappa t}$ where, N(t) is the number of cases at time t and $N_0 = N(0)$. The growth rate is often characterised by the so called doubling time, the time taken for the number of cases to double. It is given by $t_d = \ln(2)/\kappa$. Thus,

$$log_2(N(t)) = t/t_d + log_2 N_0 \tag{1}$$

1. We extract the doubling time by fitting the plot of $log_2(N(t))$ against t to a straight line. The results are shown in figure 4

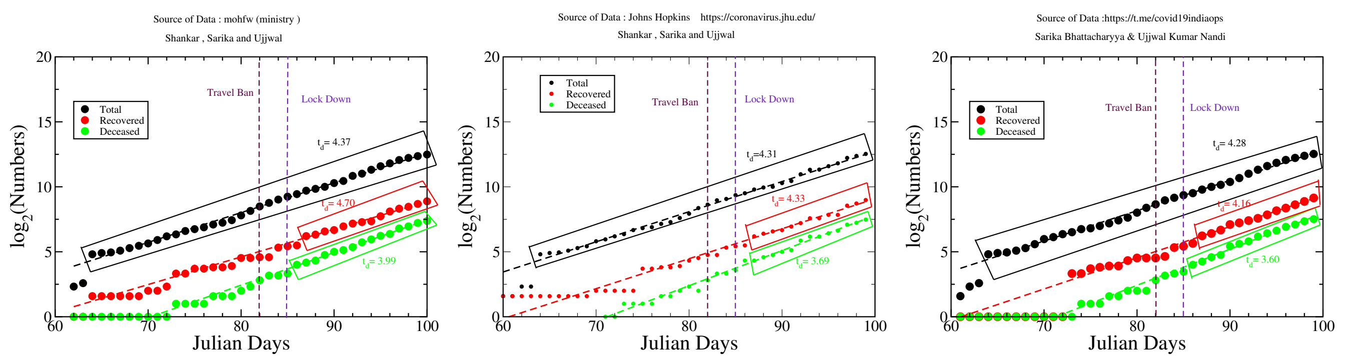

Figure 4: The confirmed, recovered and deceased cases plotted on a semilog scale for the period March 1 to April 10, 2020. The dotted lines are the best fit straight lines. The boxes indicate the points used for fitting. The doubling times are given above each line.

2. We have then extracted the doubling times for six periods, March 15-19, March 20-24, March 25-29, March 30-April 3, April 4-8 and April 9-12 of the confirmed cases from three datasets. The corresponding results are shown in figure 5.

3. The confirmed cases are divided into those with a travel history and those without, using labelling of cases done in dataset D3. The plot of the three time series, with travel history, without travel history and the total are shown in the upper panel of Figure 6.

   The doubling times for periods where the plot was quite linear was estimated and the results are shown in the lower panel of Figure 6.

2.3.3 Interpretation and discussion

1. As can be seen from figure 4 the growth rates for all the three variables has remained more of less constant from the middle of March to the first week of April.

2. The more detailed examination in figure 5 shows that there are discernible fluctuations of the growth rate of confirmed cases about the average behaviour.

3. The doubling times of the three data sets for the same time period are different, in several cases there is no overlap between the error-bars. This probably reflects the variations in the data sets discussed above.

   However, the qualitative trends of all three are similar. So we attempt provide a tentative qualitative interpretation of these fluctuations based on figure 6.

   (a) As seen from figure 6, the growth rate of the confirmed cases from around March 5 till around March 22 is largely controlled by the cases with a travel history. Assuming that most of these cases were infected when they were abroad, the growth rate during this period does not reflect the rate of transmission in India. It is reflects the rate at which the infection was injected into India.

   Why then are the numbers exponentiating during this period ? One possible answer is that since the cases outside India was exponetiating, so were the numbers that leaked into India.

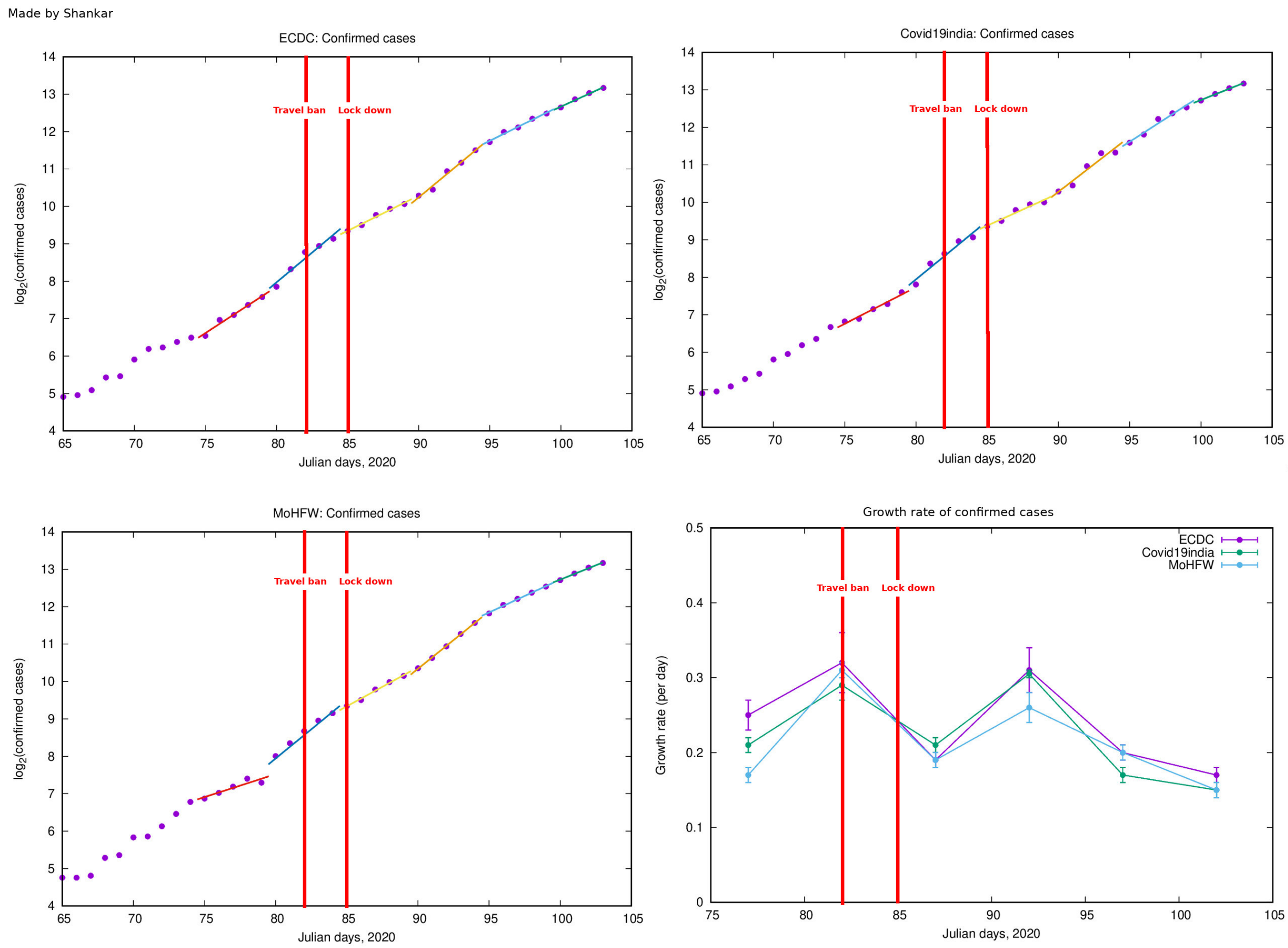

Figure 5: The three datasets with the best fit lines for the periods detailed in the main text are shown in top two and bottom left panels. The growth rates along with uncertainties for these periods are shown in the bottom right panel.

(b) From around March 19, the growth rate of cases without a travel history sharply increased, however examining the overall shape of the curve, it may be a consequence of a problem with the data. In any case the growth rate of the rest remained in the range 0.3/day to 0.4/day from March 22 to around April 4. The growth rate of the injected cases fell sharply after March 22, when the travel ban was imposed.

We interpret the decrease in the growth rate in the period March 23-30 compared to the period March 19-22 due to the sharp fall in the the growth rate of the injected cases. We do not attribute it to the lock down because we feel that it is very unlikely that the effects would be so immediate.

(c) The number of the cases without travel history exceed the number with a travel history around March 28 and started controlling the growth rate of the total soon after. We interpret the increase in the growth rate during the period March 31-April 5 as compared to March 23-30 due to this. Namely, the numbers of the cases without travel history became dominant and since their growth rate was higher, it increased the total growth rate.

(d) In the period April 3-9 the total growth rate is completely controlled by the growth rate of the cases with no travel history. The decrease in the growth rate during this period may be due to the lock down.

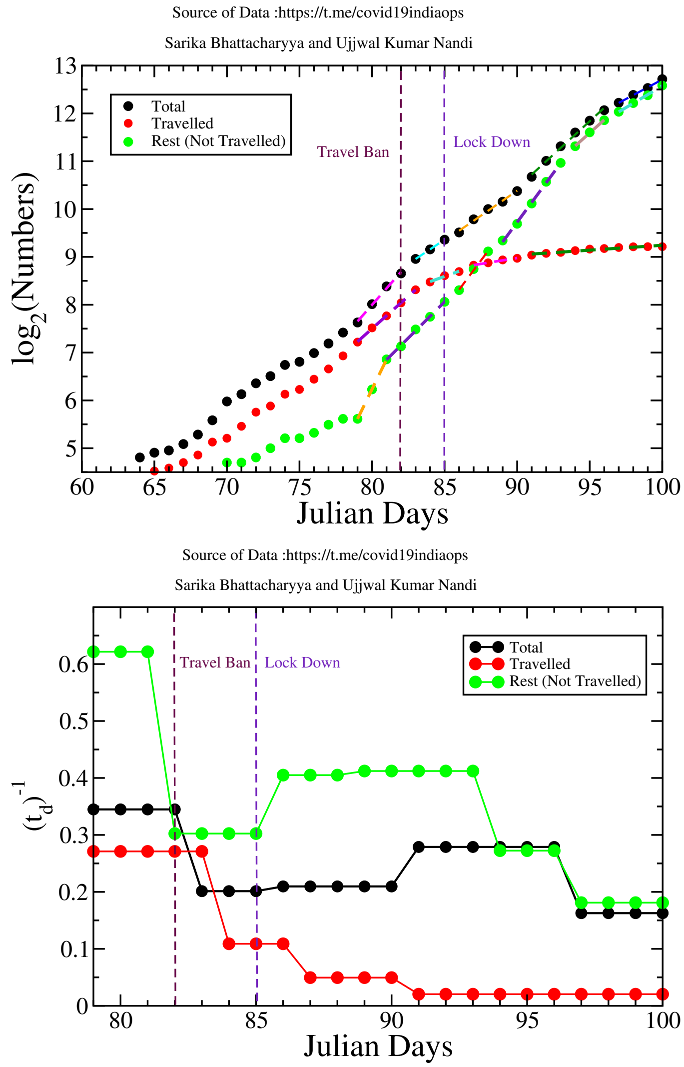

Figure 6: (Top Panel) Plot of $\log_2$(Numbers) vs days. The "Total" confirmed cases in India (black), those with a foreign travel history "Travelled" (red) and those without a travel history "Rest" (green).(Bottom Panel) The t_d^{-1} values obtained from the three plots in the left panel.

3 State level analysis

From the previous analysis it is clear that if we want to study the growth of Covid-19 in India then we should work with the cases with no foreign travel history. We analyse such data in D3 (Covid19india) for several states in this section.

3.1 Variation in growth rates

3.1.1 Question

How do the growth rates of the cases with no foreign travel history vary among different states and what has been their time evolution ?

3.1.2 Methodology and results

1. In this section we estimate the time evolution of t_d, of the population with no foreign travel history in Delhi, Kerala, Madhya Pradesh, Maharashtra, Rajasthan, Tamil Nadu, Telangana and Uttar Pradesh. These are some of the states with the largest numbers of confirmed cases. The plot of data for a selection of these states is shown in Fig.7. None of the plots show a constant t_d value and as expected it evolves with time.

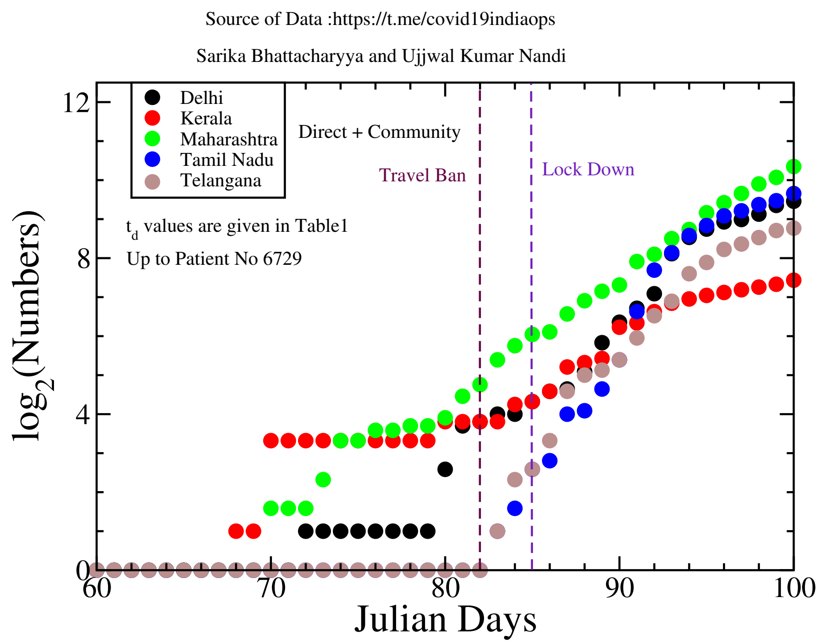

Figure 7: *$log_2(Number)$ vs Days form 1st of March 2020 for four major states. The value of t_d is lowest for Maharashtra.*

2. We have fitted straight lines to the periods of each such curve, which appear be linear, for each state and have obtained a t_d value for that period. The t_d values thus obtained are given in Table 1.

3.1.3 Discussion

1. The current doubling times range from 11.3 days (Kerala) to 4.5 days (Maharashtra). In April, the doubling times have been increasing in all the states. However, Delhi shows a slight dip in the second week of April.

2. The current doubling time for the national level data is about 5 days, which is less than that of four of the five major states. This is because the national growth rate is a weighted average of the growth rates of the individual states, the weights being the number of cases,

$$N(t) = \sum_n N_n(t) \;\Rightarrow\; \kappa(t) = \frac{1}{N(t)} \sum_n \kappa_n(t) N_n(t)$$

. Thus the national growth rate is probably being controlled by Maharashtra.

3.2 Community transmission

3.2.1 Question

What is the extent of community transmission in the different states ?

3.2.2 Methodology and results

We separate the population with no travel history into two parts, using cluster analysis as reported by D3 – people who have been contact traced to an infected individual, we label as "Direct" and people for whom contact tracing was not possible, we label as "Community". We take the ratio Community/Direct to be a measure of the extent of community transmission. We plot this ratio for the five states in Figure 8. We have also counted that on April 9, 8 of the 17 states that had reported positive cases till then had this ratio greater than 10.

3.2.3 Discussion

1. On April 9, in eight out of seventeen affected states, almost 50%, this ratio is greater than 10. This indicates that community transmission was well established in these states by then.

Table 1: Double time t_d for few major states date wise

Date	DL	KL	MP	MH	RJ	TN	TL	UP
14/04/2020	4.53	43.21	4.95	4.45	5.49	11.93	6.75	4.30
13/04/2020	4.53	43.21	4.95	4.45	5.49	11.93	6.75	4.30
12/04/2020	4.53	43.21	4.95	4.45	5.49	11.93	6.75	4.30
11/04/2020	3.44	16.23	5.40	5.21	3.09	8.95	19.57	13.62
10/04/2020	3.44	16.23	5.40	5.21	3.09	8.95	19.57	13.62
09/04/2020	6.05	11.28	5.40	4.49	3.09	7.04	7.14	13.62
08/04/2020	6.05	11.28	4.70	4.49	4.87	7.04	7.14	7.53
07/04/2020	6.05	11.28	4.70	4.49	4.87	7.04	7.14	7.53
06/04/2020	8.12	9.82	4.70	4.07	4.87	5.34	7.14	7.53
05/04/2020	8.12	9.82	5.85	4.07	4.87	5.34	3.20	7.53
04/04/2020	8.12	9.82	5.85	4.07	3.07	5.34	3.20	2.05
03/04/2020	1.39	9.82	5.85	3.48	3.07	2.27	3.20	2.05
02/04/2020	1.39	9.82		3.48	3.07	2.27	2.12	5.91
01/04/2020	1.39	4.24	1.68	3.48	3.07	2.27	2.12	5.91
31/03/2020	2.27	4.24	1.68	3.48	4.81	1.19	2.12	5.91
30/03/2020	2.27	4.24	1.68	4.04	4.81	1.19	3.92	5.91
29/03/2020	2.27	9.22		4.04	7.08	1.19	3.92	2.49
28/03/2020	1.68	9.22	1.79	4.04	7.08	1.19	3.92	2.49
27/03/2020	1.68	9.22	1.79	4.04	7.08	1.38	1.38	3.56
26/03/2020	6.96	5.93	1.79	5.65	3.88	1.38	1.38	3.56
25/03/2020	6.96	5.93	1.79	5.65	3.88	1.38	1.38	3.56
24/03/2020	6.96	5.93		5.65		1.38	1.38	
23/03/2020				2.1	4.12	1.38		
22/03/2020				2.1	4.12			
21/03/2020				2.1	4.12			
20/03/2020				2.1				

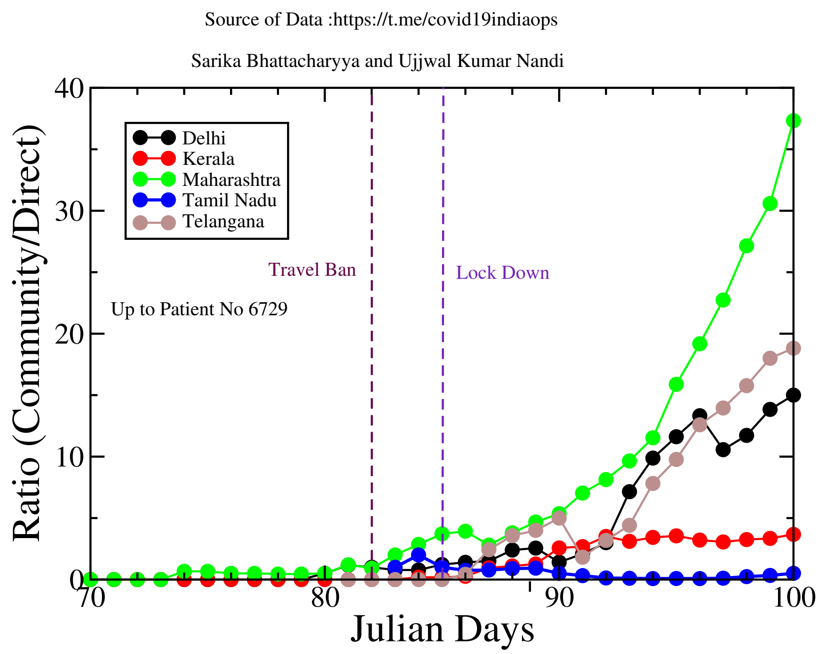

Figure 8: *The ratio of community number to the direct number is plotted. The higher the value of this ratio the more is the community spread.*

2. Maharashtra has the highest value of this ratio which is growing quite rapidly with time.
3. The Community/Direct ratio for Kerala has plateaued to a value below 5 in the same time period (around 30th March) when it's t_d value more than doubled (increased from 4.24 to 9.82).
4. Although Tamil Nadu has a large number of cases, this ratio is very small for the state. This probably will make it easier for this state to contain the infection.

3.3 Maharashtra

3.3.1 Question

The question that we ask is why the number of cases is growing so rapidly in Maharashtra compared to the other states ? Is it because of large community transmission or because there are more tests being done in this state?

3.3.2 Methodology and results

1. We looked at the data from the two most affected cities in Maharashtra, Mumbai and Pune. In Fig.9(left panel) we plot the $log_2(number)$ vs. days for Mumbai and Pune. In the right panel, the community spreading ratio for these two cites is shown. The most affected city in Kerala, Kasaragod has been added for comparison.
2. In the left panel of Fig.10 we plot the test data taken till 7th of April for a few states. In right panel we plot the number of tests per million population. Here population is taken from (http://statisticstimes.com/demographics/population-of-indian-states.php) and this data provides the population till 2018.

3.3.3 Discussion

1. When we compare the t_d values over the last week for Mumbai and Pune, we find that they are quite similar. However the community ratio shows that compared to Pune, Mumbai has a huge community transmission. Both cities have a large ratio compared to the most affected city in Kerala, Kasaragod.
2. We find that indeed till 7th of April, MH has done the maximum number of tests in India. However, compared to MH, Kerala did more number of tests per million.
3. The analysis shows that the large number of cases in MH cannot be only attributed to the carrying out of large number of tests. Community spread appears to be one of the biggest reason for this rise in number.

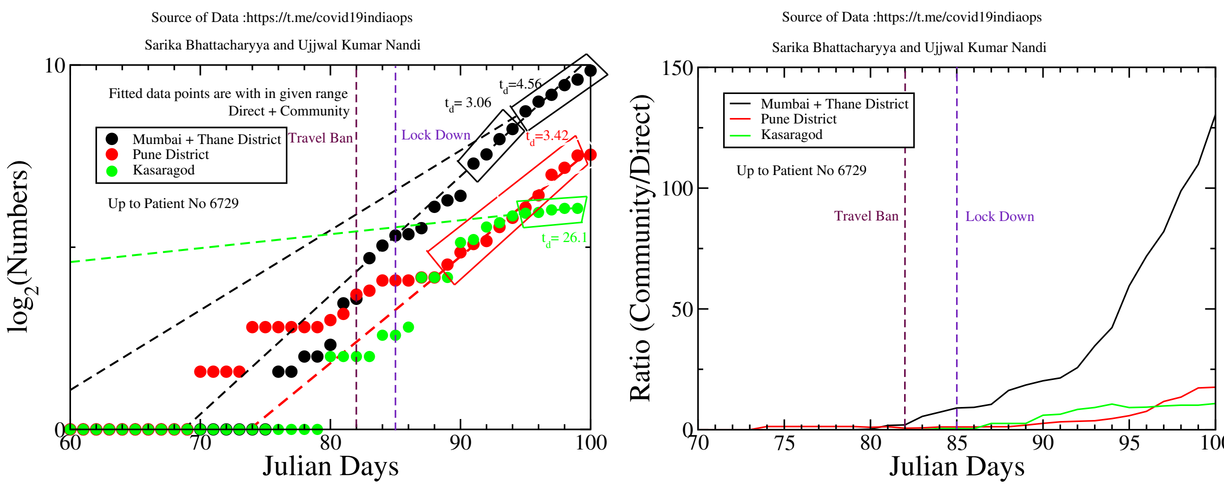

Figure 9: *(Left Panel)*$log_2(Number)$ *vs Days form 1st of March 2020 for two big cities in Maharashtra (Mumbai and Pune). The value of* t_d *is lowest for Maharashtra. For comparison we also plot the same for the most affected city in Kerala, Kasaragod.(Right Panel) The ratio of community number to the direct number is plotted. The higher the value of this ratio the more is the community. spread. For comparison we also plot the same for the most affected city in Kerala, Kasaragod.*

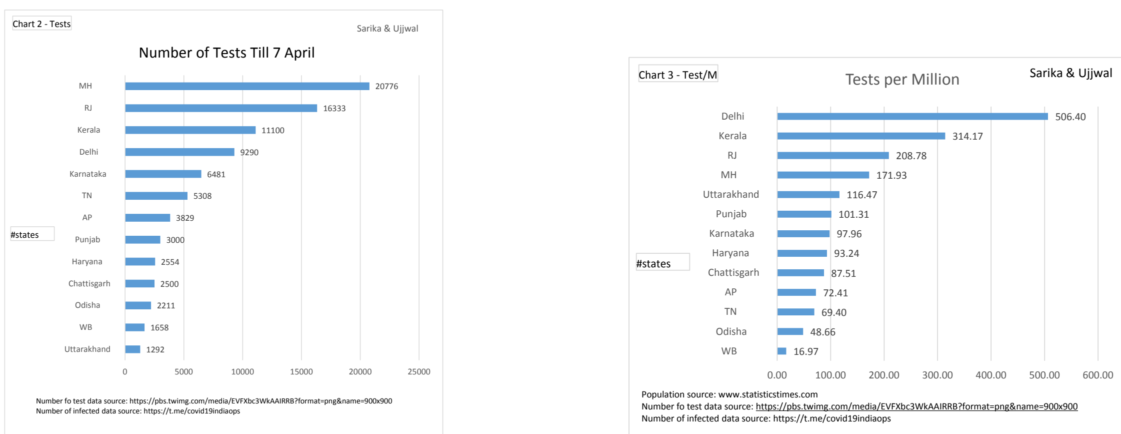

Figure 10: *(Left Panel)Total number of tests in some states till 7th of April 2020. MH has done the maximum number of tests. (Right Panel) The tests done per million people in that state.*

4. Since MH appears to dominate the growth rate in India it definitely needs a change in its strategy for management of the spread of the disease.

5. Kerala seems to be a good model state in containing the infection. Thus other states should follow the method implemented in Kerala.