CS85A: Data Mining Assignment #1

Aditya Jain
20111004
adityaj20@iitk.ac.in

Indian Institute of Technology, Kanpur— September 25, 2020

1 Introduction

India is facing an unprecedented crisis due to the novel coronavirus, first reported in Wuhan, China in December 2019. In India COVID was first reported on January 30, 2020, in Kerala and it upsurged very quickly to all states in the country. The total number of confirmed cases are more than 54 Lakhs to date (September 20, 2020). It has disrupted the lives of 1.3 billion Indians, first with a complete lockdown in the country from March 24 - May 31, 2020, then various aspects as work from home, etc.

In this project, I have tried to analyse the growth of cases in each district for every week, month and overall from **March 15 - September 5, 2020.** There were total of 112 cases before March 15 and 41,10,839 on September 5, 2020. Districts are classed as hotspot and coldspot per week, month, and overall concerning their neighbours and in overall state.

1.1 Data collection

Data for this study is collected from covid19india.org. Following csv files are incorporated for data processing to generate the results:

1. raw_data1.csv - It contains patient wise data till April 19.
2. raw_data2.csv - It contains patient wise data from April 20 to April 26.
3. districts.csv - District wise timeseries of Confirmed, Recovered and Deceased numbers from April 26 to Sep 6 2020.
4. neighbor-districts-modified.json - List of neighbours of each district of India

1.2 Data Pre-processing

- Due to errors in district-wise data of state Assam, Telangana, Goa, Delhi and city Mumbai I have merged all their districts as one single district.
- Cases of districts that didn't map to the neighbour-district file and the ones titled as "Unknown" are removed from the dataset.

A total of 41,10,839 cases are recorded in the input files from covid19.org for the period March 15 to September 5, 2020. After complete processing total of 40,50,336 cases for 645 different districts are obtained.

2 Observations

2.1 Total Cases as of Sep 5, 2020

Following graph shows the total number of confirmed COVID cases for each district in India.

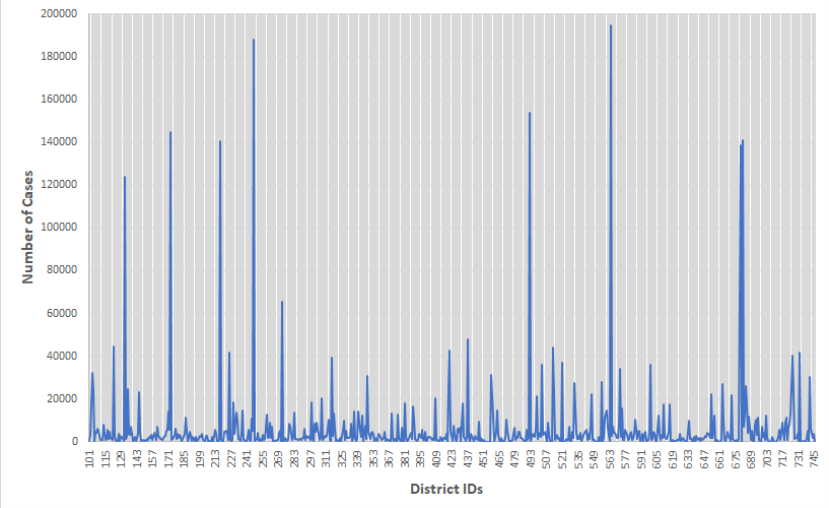


Figure 1: Total recorded cases for each district.

- There are nine districts with more than 1 lakh cases in which Pune, Delhi, Mumbai and Bengaluru are on top.
- There are 17 districts more than 30 thousand cases.
- Rest 619 districts have fewer cases.

These results show that major cases are restricted to some districts rather than equal distribution. I haven't included population per capita statistics in our analysis which may be the reason our results are showing some highly populated districts as hotspots.

2.2 COVID-19 entrance in India

The COVID-19 cases appeared in March 2020 were related to people who have been evacuated or have arrived from COVID-19–affected countries. There were 632 such foreign cases out of 1013 total cases to date March 23, 2020, from which government suspended international travel till July 31, 2020.

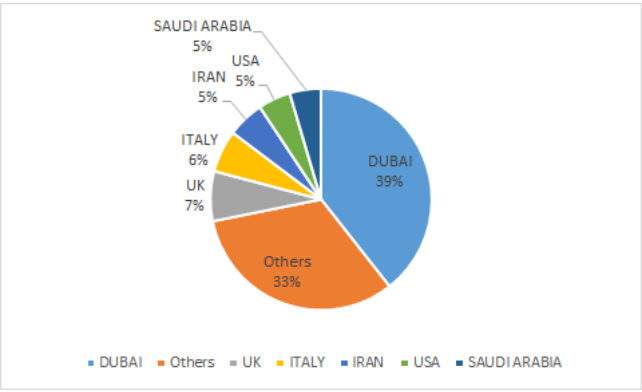


Figure 2: Cases arrived from foreign countries before international travel ban.

Travellers from Italy, Dubai, US, and the UK comprise 61% of these cases, and Kerala was the most affected by these foreign cases with a count of 221.
28 out of 36 states have recorded at least a case with foreign travel history. These foreign cases were the start of the pandemic in the country.

2.3 Month-wise spread

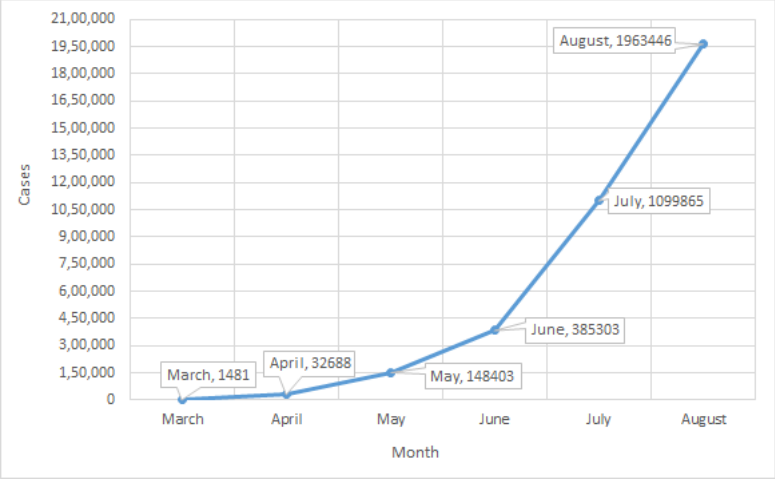


Figure 3: Curve representing number of new cases generated every month

The number of new cases reported each month is approximately 2.39 times the cases generated in the previous month except for April. The number of new cases in April was 4.5 times the number reported in March. So data suggests that:

- India was in stage 1 (disease through people with travel history) of COVID-19 in March with only 169 affected districts.
- Stage 2 (local transmission - friends and family of people with travel history) started in April with 422 affected districts.
- Stage 3 (community transmission - when the source for the virus cannot be traced) started in May with 592 affected districts.
- 641 districts out of 645 districts had reported a case of COVID-19 till June 2020. All 645 districts were affected by August, 2020.

2.4 Week-wise spread

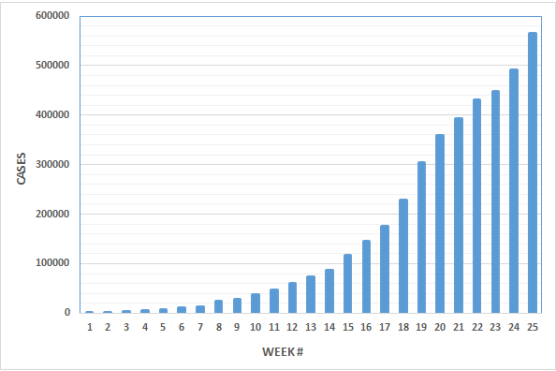


Figure 4: Cases spread over the weeks

I have discretized the data set into weeks such that March 15 - March 21, 2020, is week 1 and so on. Above graph represents the number of new cases reported each week from March 15 to September 5, 2020. Graph shows exponential growth in cases over weeks.

2.5 Hotspots Analysis

2.5.1 Hotspot among neighbouring districts

I have categorized the districts into a hotspot if the number of cases recorded in the districts is greater than the sum of the mean and standard deviation of its neighbour districts cases. Following is the graph showing number of hotspot districts in each month:

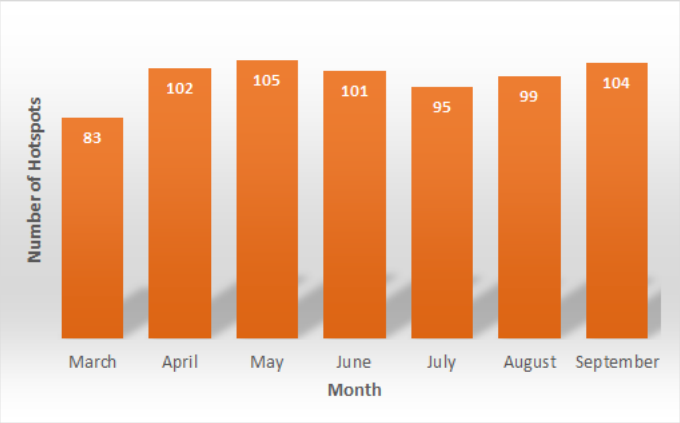


Figure 5: Number of hotspot districts among their neighbors over months

A district is hotspot if the difference between the number of cases incurred in it is very large than the cases incurred in its neighbouring districts. Analysis shows that on an average a district remains a hotspot for around 3-4 months only. There are only 21 such districts which remained hotspots for all seven months of the analysis. This means that once a district becomes a hotspot, there is an increase in growth rate of COVID cases in its neighbouring districts, reducing the difference between the number of cases in hotspot and its neighbouring districts.
So we can speculate that the infection spreads from a hotspot district to its neighbours within a month period, as very few districts remain hotspots among its niehgbours as we move across the timeline.

2.5.2 Hotspot within a State

I have categorized the districts into a hotspot if the number of cases recorded in the districts is greater than the sum of the mean and standard deviation of its state districts cases. Following is the graph showing number of hotspot districts in each month:

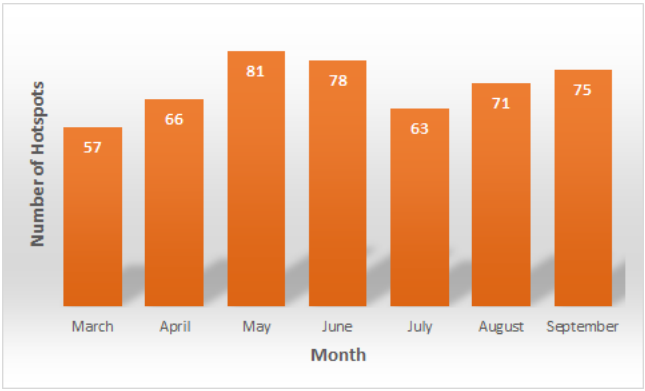
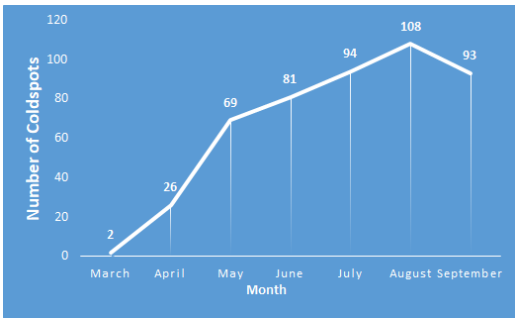


Figure 6: Number of State hotspot districts

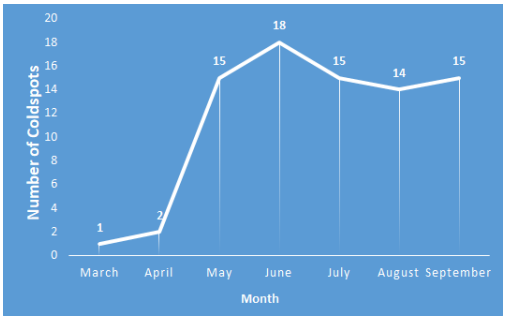
Since there was a significant hike in the number of hotspots in May, it shows that in May and June most cases were restricted to the main hub cities of the state. This conclusion is also backed by the fact that inter-district travel was highly restricted due to nation-wide lockdown placed by the government until June 2020. Till July most districts of the state started recording significant cases, and hence the number of hotspots get normalized to normal range from July.
Analysis shows that only 18 districts remained hotspots in all the months, so we may conclude that virus migrated in large numbers from these hubs to other districts of the state in July when the lockdown was also removed.

2.6 Coldspots Analysis

I have categorized the districts into a coldspot if the number of cases recorded in the districts is smaller than the subtraction of the mean and standard deviation of its neighbour districts (for neighbour analysis) and state districts (for state analysis) cases. Following is the graph showing number of coldspots districts in each month for both the ases:



(a) Number of State coldspot districts



(b) Number of coldspot districts among their neighbors

We observed that the number of coldspots have increased every month for both the cases (State and Neigbour analysis). This increase of coldspots also shows that a large number of patients are restricted to some districts. Cases are rising at a tremendous rate in these hubs with respect to the other districts in the state, making their neighbours classify as coldspots. These hubs are the districts with a large population per capita density.
The number of coldspots will keep on increasing to a point and then will gradually fall when the majority of the population in these hubs are recovered from the disease.

3 Conclusion

COVID cases in India are rising day by day and don't seem to stall as the infection has just penetrated 0.3% of the population. A large number of cases are restricted to some districts only their growth rate is tremendious as compared to the districts around them.

We saw that number of hotspots are constant among the months, and the number of coldspots is increasing every month. This pattern will follow due to different population per capita in different districts.