Graph Frequency Analysis of COVID-19

Sreenya Chitluri - 2020102065
Smruti Biswal - 2020112011

August 2021

Contents

1 Abstract

1.1 Aim

1. The aim of the project was to perform a spatio-temporal analysis on the spread of COVID-19 in 100 different districts of India.

2. To study the outbreak during the 'second wave' to speculate plausible solutions.

1.2 Motivation behind the approach

All the pre-existing work in this area concentrates only on a single varying parameter. However, this study links two different yet important variables and attempts to establish a link between the two.

1.3 Parameters

1. Time: It uses the data related to number of COVID cases across 100 different districts in India spanning a period of 80 days starting from April 1$^{\text{st}}$ 2021 to June 19$^{\text{th}}$ 2021.

2. Commute Flow: It uses the data related to the commute flow between districts. Since this data is not openly available for India, we used an alternative parameter, distance between the districts, which contributes to the same. Districts that are closer to each other are expected to have more population mobility between them.

Concepts of graph signal processing were used to perform the analysis.

2 Important concepts and short discussion about the existing method

2.1 Concepts

1. **Graph and signals:** We build a graph with the nodes being the 100 districts chosen, edges being the connection between two districts and weight of the edges being the distance between them. Each of the node has a signal associated with it which corresponds to daily confirmed COVID-19 cases over the 80 days.

$$x = [x_1, x_2, ..., x_{100}]^T \text{ where } x_i \text{ represents the signal for the } i^{th} \text{ node}$$
$$\mathcal{G}(\mathcal{V}, \varepsilon, \mathbf{W}) \text{ is the graph where } \varepsilon \subseteq \mathcal{V} \times \mathcal{V}$$
$$\mathbf{W}_{ij} \rightarrow \text{distance between } \mathcal{V}_i \text{ and } \mathcal{V}_j \text{ where } \mathcal{V}_x \text{ denotes the } x^{th} \text{ node.}$$

2. **Graph Fourier Transform:** Graph Fourier Transform changes the domain from time to spectral frequency in the graph signals.

$$\mathbf{L} = diag(\mathbf{W1}) - \mathbf{W} \text{ where } \mathbf{W1} \text{ is the degree matrix of the graph.}$$
$$\mathbf{L} = \mathbf{V} \wedge \mathbf{V}^T$$
$$\mathbf{V} \rightarrow \text{eigenvectors of } \mathbf{L}$$
$$\wedge \rightarrow \text{eigenvalues of } \mathbf{L}$$
$$\text{GFT} \rightarrow \tilde{x} = \mathbf{V}^T x$$
$$\text{iGFT} \rightarrow x = \mathbf{V}\tilde{x}$$

3. **Graph Filtering:** We construct a LPF and a HPF by isolating the first and the last $\frac{1}{5}^{\text{th}}$ of the eigenvalues and their corresponding eigenvectors respectively.
   LPF construction:
   $$x_L = \mathbf{V}\tilde{H}_L\mathbf{V}^T x$$
   $$\text{where } \tilde{H}_L = diag(\tilde{h}_L) \text{ and } \tilde{h}_{L,n} = I\{n < N_L\},\ N_L = \frac{N}{5}$$
   A HPF can also be constructed similarly.

2.2 Existing method

The method followed for analysis:

1. We construct the graph with 100 nodes.
2. Plot the daily signals for 9 different districts.
3. Plot the signals for the same districts after normalising by population.
4. Find its GFT, pass it through the HPF and LPF, and identify the HP and LP regions of the map of India. HP regions are those which have a dominance of higher frequency components and LP regions the lower correspondingly.

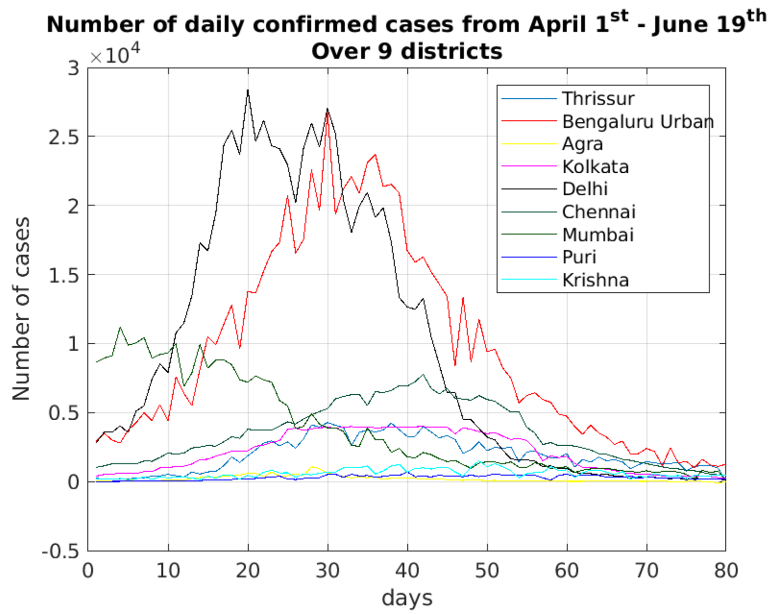
3 Proposed Methods

Instead of using data of population flow, a useful alternative metric is distance. Collect the coordinates of all the districts that are in our interest and calculate the distance between every pair using the Haversine formula. Now build the adjacency matrix of the graph. Colour code it such that the deeper hues represent smaller distances and warmer hues represent greater distances. This is an effective alternative as it is a fair assumption that the commute flow is greater when the districts are closer.

4 Simulations and Results

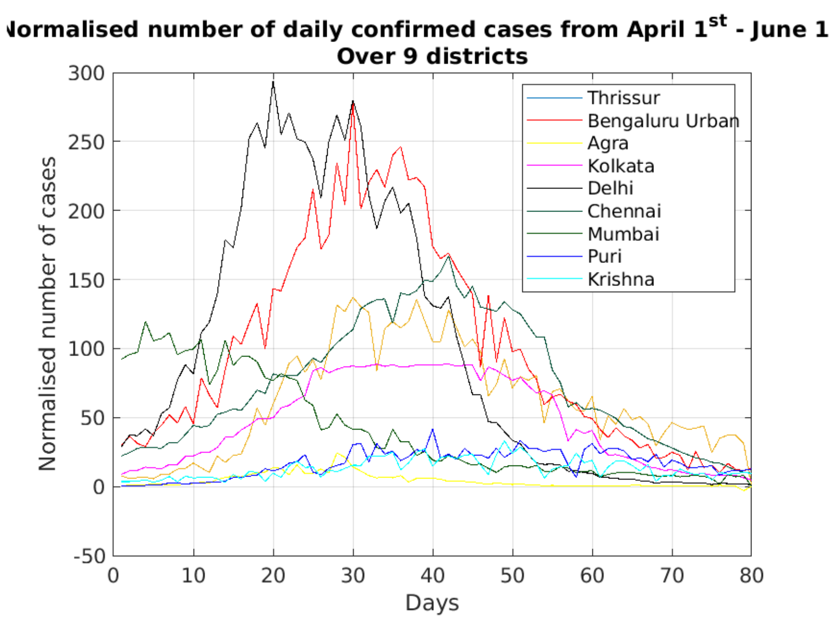
4.1 COVID Data

In the following graph we plot the number of daily cases obtained from cumulative source data. Random 9 districts were selected to represent the randomness of the cases in India. The graphs are plotted for a time period of 80 days from 1^{st} April to 19^{th} June 2021 denoting the second wave of the COVID in Indian districts.



Normalising cases:

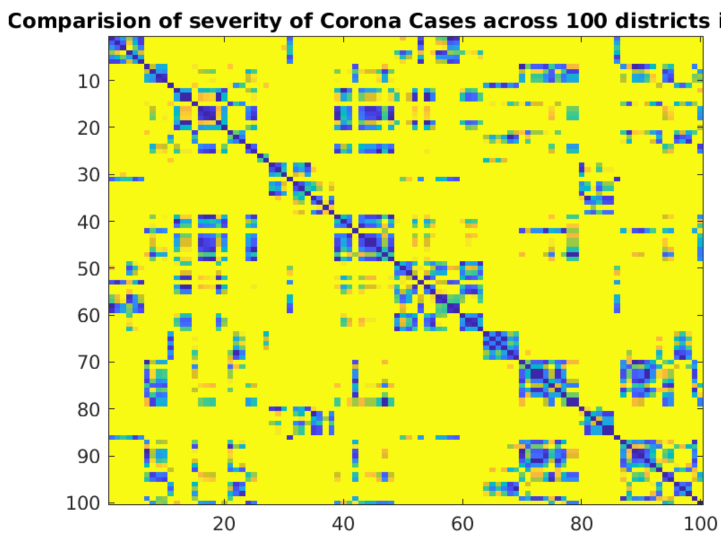
This following graph is the number of daily cases normalized over population of the districts i.e. it shows number of cases per 100k people in each district.



Importance: The normalisation of cases based on the population is necessary to understand the extent of spread of COVID in particular district during second wave. As we can see, few districts such as Kolkata, Thrissur and Agra seem to have minimal cases as compared Delhi and Bengaluru in the daily cases graph giving a false sense of the reality. But in the Normalised cases graph we can see that the spread intensity is prominent.

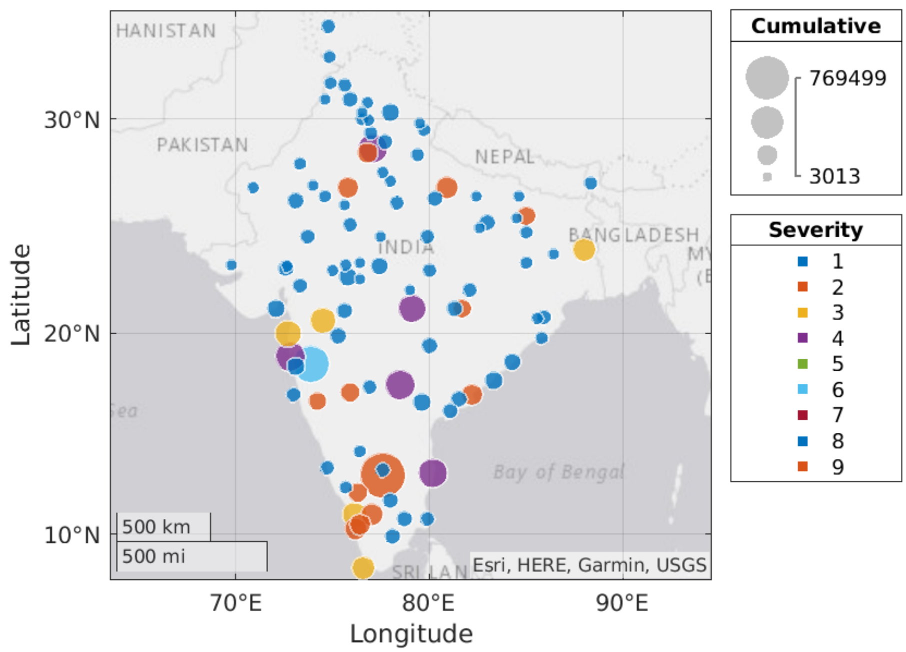
4.2 Adjacency matrix

The adjacency matrix gives the relative distance between 2 districts. The darker shades i.e. the ones that are blue end denote shorter distance between districts and hence more commute flow. The yellower regions denote more distance and hence less commute flow. This was done since the commute flow data in India is unavailable and under the assumption that if distance is less then people will travel more.

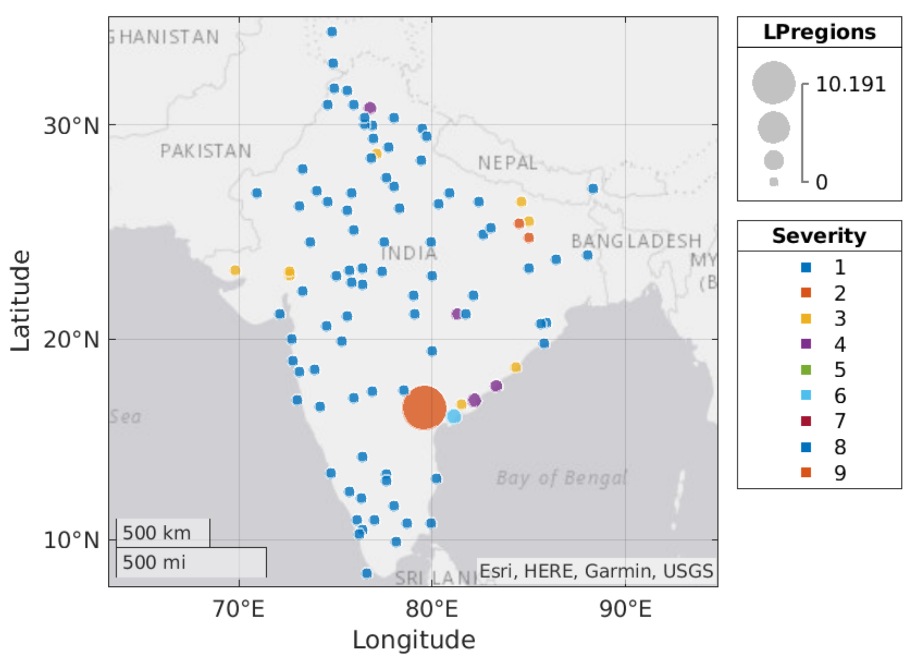


4.3 Graph Processing

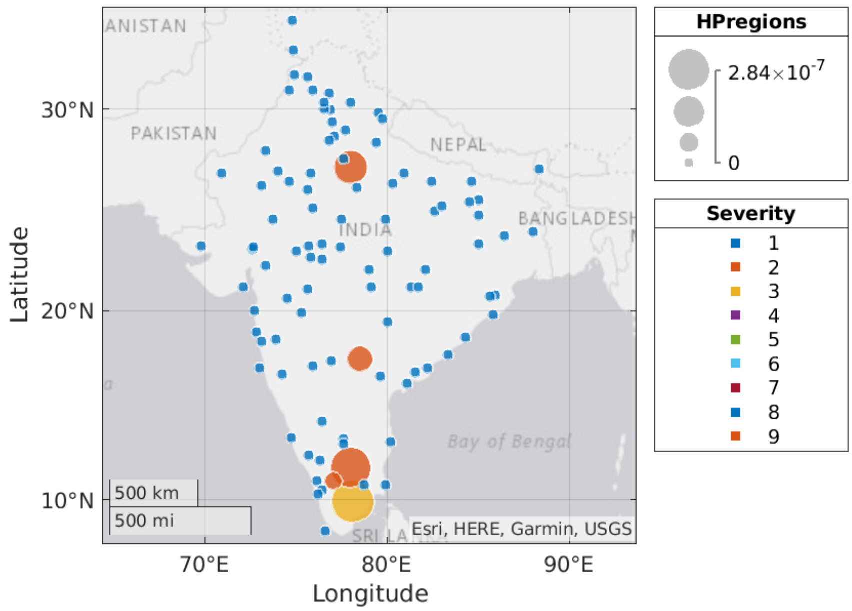
The following graph shows the cumulative cases in India for 100 different districts over 80 days. The number of cases is total number of confirmed cases from 1_{st} April to 19_{th} June 2021. The range is divided into 9 levels of severity - 1 being minimum to 9 being the most severe.The size of bubble associated is directly proportional to the number of cases of a given district.



The following graph shows the regions after being filtered through a LPF and related to the low frequencies of the signal. It shows gradual increase in the number of cases.



The following graph shows the regions after being filtered through a HPF and related to the high frequencies of the signal. It shows a sudden outbreak/spike in number of cases.



5 Conclusion

The following Conclusions can be derived:

1. The normalised number of cases give us a better measure of extent of spread of diseases.

2. The LP regions denote the lower frequencies. They tell us about the regions which have spread of Corona cases due to inter-district movement i.e. the spread is due to people moving in and out of the said district. If the region has high severity after passing it through LPF then we must constrict travel in those regions. The LP regions are generally gradual increase in cases in severity and intensity.

3. The HP regions denote the higher frequencies. They tell us that the certain region had a spike in number of cases. Thus informing of a case of sudden outbreak and there is no connection of inter-district commute flow but rather intra-district movement. The sudden spikes generally occur over a short period and disappear fast as well given we control it.

Hence these graphs give us the contagion pattern of spread of COVID-19 cases w.r.t to the commute flow over time. This study can be used to control the spread of epidemic like where must a lockdown be imposed or where we must control the activities of a given region. This can also be further used to study other pandemics and hopefully control them.

6 References

1. https://en.wikipedia.org/wiki/Laplacian_matrix

2. https://en.wikipedia.org/wiki/Haversine_formula

3. https://api.covid19india.org/csv/latest/districts.csv

4. Introduction to graph signal processing - Ljubisa Stankovic, Milos Dakovic, andErvin Sejdi

5. http://data.icrisat.org/dld/src/definition_standard.html