Pandemic Data Science: Vulnerability Assessment of States and Inter-State Rail Networks

Harshit Kumar
Department of Computer Science
IIT Gandhinagar
harshit.kumar@iitgn.ac.in

Pranshu Kumar Gond
Department of Computer Science
IIT Gandhinagar
pranshu.kumar@iitgn.ac.in

Sagar Bisen
Department of Computer Science
IIT Gandhinagar
sagar.bisen@iitgn.ac.in

Abstract

The COVID 19 pandemic has been around for over a year and it has become necessary to come up with solutions that not only tackle the problems that we have been battling but are also able to tackle those problems that may arise as the pandemic progresses. In this project, we perform Exploratory Data Analysis (EDA) on Indian Railway, Hospital, Vaccination and Demographics datasets to draw useful inferences which could help in coming up with strategies to tackle the pandemic.

1 Introduction

As of 11th May 2021, a total of 159 Million cases of SARS-CoV2 have been recorded, of which 3.3 Million were fatal. Since the pandemic was an unprecedented situation for everyone, tackling it has proven extremely difficult. But now that over a year has passed, there is a large amount of data that has been collected and compiled, enough to be analysed in detail in order to identify trends and draw inferences.

India is one of the most severely affected nations in the world, owing to its high population density. Due to its large size and dense transport networks, restricting the spread of the virus has proven to be an extremely difficult task.

The ongoing Second Wave of the virus has caused a sharp increase in the number of active cases, forcing us to reassess our counter strategies. Hospitals and hospital beds need to be managed with as high efficiency as possible. The ongoing vaccination drive is also of great importance, which needs to be properly planned and executed.

To help in all these continuous processes of planning and management, data and data analysis is of great importance. In this project, we attempt to draw useful inferences from the available compiled datasets by performing EDA, clustering and other important techniques. We also aim to assign vulnerablity scores to states and Inter-state rail networks.

Preprint. Under review.

2 Related Work

[1] used DEA analysis to assign vulnerability scores (Social and Health) to the various travel corridors between the states of Gujarat and Maharashtra. [2] used Machine Learning techniques to predict safe travel routes in New York City by assessing the crime records and patterns in the city.

assessed the outbreak clusters in India and carried out an analysis to understand the variablity in the preparedness to fight the novel Covid-19 pandemic of the various states in India. [4] used a Fuzzy Multidimensional Analysis to understand how the living standards of Brazil are directly related to the capacity to prevent Covid-19 in various households.

3 Datasets

These are the datasets that have been used in the project:

- Demographics Dataset: A district-wise dataset with 693 districts and 95 features which include total population, male and female working population and so on, all meant to describe the demographics of a given district. This dataset was then converted to state-wise dataset with 35 States/UTs and 95 features.
- Vaccination Dataset: A state-wise dataset with 35 States/UTs and 10 features which include Number of individuals vaccinated, Number of vaccines registered, Number of male population vaccinated and so on.
- Health Infrastructure Dataset: A state-wise dataset which includes the number of hospitals and hospital beds maintained by the following health organizations,
  - Hospitals maintained by the Indian Government
  - Hospitals maintained by Indian Railways
  - Hospitals maintained by the Ministry of Defence
  - Hospitals maintained by the State Government (District Hospitals, Community Health Centres, etc.)
- Indian Railway Network Dataset: This dataset includes the list of trains active in the Indian Railway Network. Total Number of trains included in the dataset is 5199. Total number of stations are 422. However, due to present of null values and other constraints, we only use a total of 432 trains and 292 stations.
- Case Count Dataset: A district-wise dataset with 693 districs and features that include the number of Active cases, Recovered cases and Deaths. This dataset was converted to a state-wise dataset with 35 states/UTs.

4 DBSCAN

Hierarchical and Partition based Clustering Techniques are the most commonly used clustering techniques. However, they fail to capture arbitrarily shaped clusters or detect outliers. When dealing with abnormally shaped clusters, density based clustering techniques are far more efficient.
Density Based Spatial Clustering of Applications with Noise (DBSCAN) [5] is one of the most simple density based clustering techniques. There are 2 parameters in DBSCAN : `eps` and `minPts`. These 2 parameters determine whether a point will be a *Core Point*, *Border Point* or *Outlier*.

- *Core Point* : A point is said to be a core point if there are at least `minPts` points within a distance of `eps` from the point.
- *Border Point* : A point is said to be a border point if there are less than `minPts` points within a distance of `eps` but at least one of them is a core point.
- *Outlier* : A point is said to be an outlier if there are less than `minPts` points within a distance of `eps` and none of them is a core point.

Based on this classification principle, this is the algorithm :

- A starting point is chosen at random.

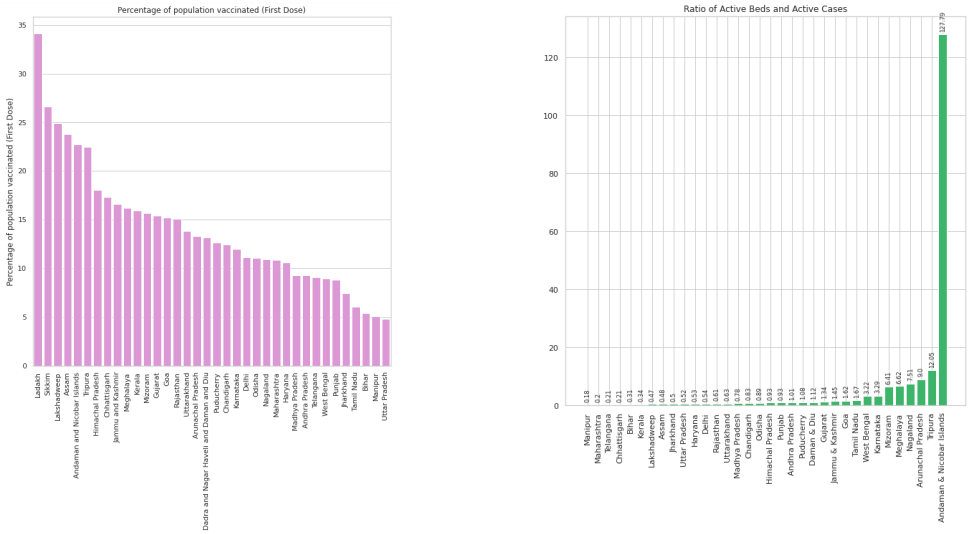

Figure 1: The first figure shows the percentage of population vaccinated wiith the first dose in each state and the other figure shows the ratio of Total Beds and Active cases for each state

- If it is not a core point, it is marked as noise.
- If it is a core point, cluster formation starts. Point marking and addition to cluster is done recursively.
- If there are unvisited points remaining, another starting point is chosen at random.

The smallest cluster possible would be a single core point and `minPts` border points. One drawback of DBSCAN is that sparse clusters are not detected by the algorithm. Also, parameters must be in an optimum range. Extreme cases could be either all points being placed in a single cluster or all points being classified as outliers.

5 Method

5.1 EDA

We performed EDA or Exploratory Data Analysis on the datasets that we had gathered. This was done to understand the datasets better. Using the analysis, we came to understand the important features that we can consider for the calculation of the vulnerability metrics. Figure 1 show the percentage of population vaccinated (first dose) in a given state and the ratio of total hospital beds and active cases in a given state.

Feature	Correlation to the Number of Active Cases
Marginal workers population (others)	0.801850
Literate population	0.761175
Number of households	0.735023
Female Population	0.697729
Male Population	0.691201

Table 1: The Correlation Matrix showing the strong correlations of demographic features to the number of Active Cases

Table 1 shows the the various features from the demographics data that had the maximum correlation with the number of active cases in a state. These features are then used to calculate the state demographic index (Section 6.2.1).

5.2 Vulnerability Metrics

We have defined two vulnerability metrics, i) The State Vulnerability Score ii) The Rail line Vulnerability Score

5.2.1 The State Vulnerability Score

The state vulnerability metric represents how severely the state has been affected due to the ongoing pandemic. This metric is calculated using the inferences gathered after performing EDA on the compiled datasets (Section 4.1).
We concluded that the *percentage of population vaccinated* (with the first dose) in a given state can be considered a valid factor in determining how effective the vaccination protocols have been enforced in that state. Similarly, The *ratio of total available beds (in a state) and the number of active cases* gives an estimate about how stable the medical infrastructure is in a state. Further, we calculated a *state demographic index* to incorporate the features of the state demographics that had high correlations with the number of active cases in the state. Both these factors should have a negative correlation to the state's vulnerability.

The state demographic index was calculated as follows,
Firstly, we selected the features with high correlation to the number of active cases. These features are, i) Marginal Workers Population (People with non industrial and non agricultural jobs) ii) The literate population iii) Female Population iv) Male Population v) Number of Households
These features were normalised using the formula given below

$$y_{norm} = \frac{y_i - y_{min}}{y_{max} - y_{min}} \tag{1}$$

Where y_{norm} is the normalised feature, y_i is the non normalised feature, y_{min} is the minimum value of that feature in the dataset and y_max is the maximum value of the feature in the dataset.
Let δ_d denore the State Demographic Index then,

$\delta_d =$ Marginal Other Workers Population Person + Literate Population + Female Population + Male Population + Number of Households

Let FV be the fraction of population vaccinated, let r_{beds} be the ratio of Total beds and Active Cases and let SV be the state vulnerability.

$$SV = \delta_d + (FV \times r_{beds})^{-1} \tag{2}$$

5.2.2 The Rail line Vulnerability Score

For calculating the Rail Line Vulnerability Score, we consider the duration of the train journey and the state vulnerabilities of the state from where the train journey begins and the state at which the journey ends.
For calculating the Rail Line we assume the following,

1. Greater the duration of the journey, higher will be the risk of getting infected.
2. The State Vulnerability Score should have a positive correlation to the Rail line Vulnerability Score.

We use equation (1) to normalise the duration of the train journey.

let t_{norm} be the normalised duration of the journey, let SV_{start} be the State Vulnerability Score of the state from where the train starts it's journey, let SV_{end} be the State Vulnerability Score of the state where the train ends it's journey. Finally, let RV denote the railway network vulnerability

We compute RV as,

$$RV = \frac{SV_{start} + SV_{end}}{2} \times t_{norm} \tag{3}$$

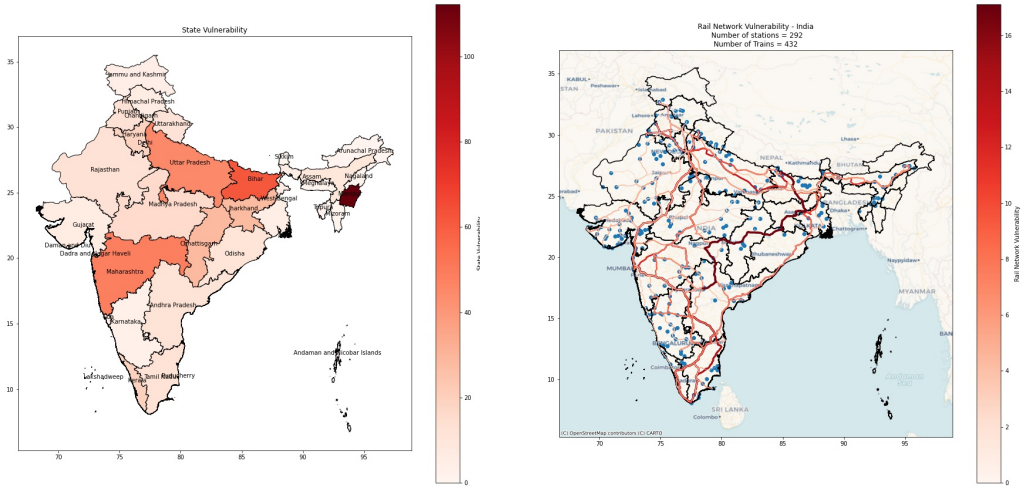

Figure 2: Chloropleth Plot of the state of India based on their calculated State Vulnerability Score and chloropleth plot of the railway lines based on the calculated Rail line Vulnerability Score

5.3 DBSCAN Clustering

DBSCAN clustering was performed on the District-wise Datasets for Demographics and Vaccination. Steps involved were :

1. Reading data from `.csv` file into `pandas` dataframe
2. Dropping ineffectual columns
3. Scaling the values using the `StandardScaler` method in `sklearn.preprocessing` library
4. Performing PCA on the data to transform it into 2 dimensions using `PCA(n_components=2)` method from `sklearn.decomposition` library
5. Running `DBSCAN` from `sklearn.cluster` library to classify points into appropriate clusters
6. Plotting the final cluster map using `plotly.express`

6 Results and Discussion

6.1 Vulnerability Metrics

After calculating the State Vulnerability Score for each state using equation (2), we plotted a chloropleth plot (Figure 2) using the generated score. We observe the following,

1. Manipur comes out to be the state with the highest State Vulnerability Score. This is because the percentage of population vaccinated is quite less in Manipur (only 5.031%). Also, the ratio of total beds and the active cases in Maniput is the lowest (0.177)
2. Some of the other states with high State Vulnerability Scores are Bihar, Uttar Pradesh, Maharashtra and Chhattisgarh.

After calculating the Rail line Vulnerability Score of all the rail networks we had, we plotted a chloropleth plot shown in figure 2. Following inferences were drawn,

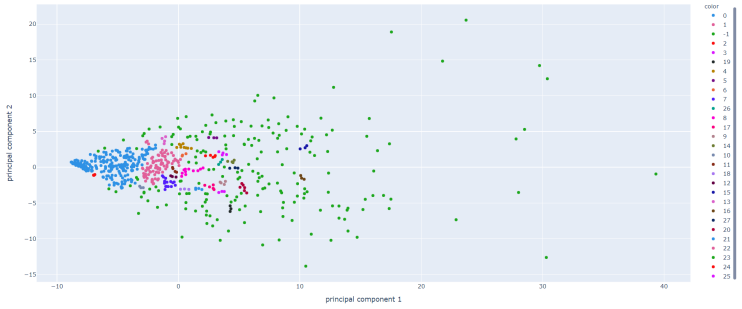

Figure 3: Clustering Plot of Demographics Dataset

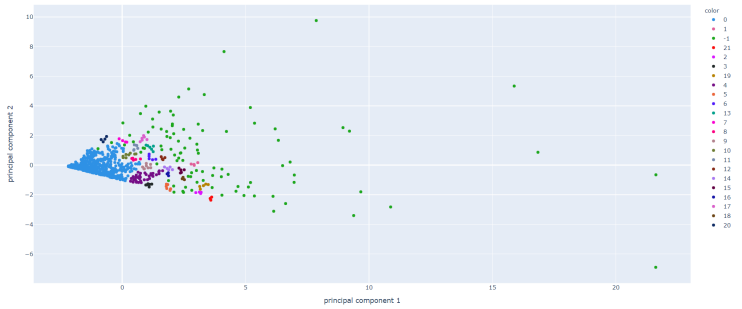

Figure 4: Clustering Plot of Vaccination Dataset

1. As expected, the trains which had the higher vulnerabilities were the ones travelling from a state with high State Vulnerability Score or were travelling to a state with high Vulnerability Score or both.
2. An interesting thing to note is how the duration of the journey plays a major role in determining the Rail line Vulnerability Score. Trains with higher duration of journey tends to have higher vulnerabilities compared to trains travelling to or from states with high State Vulnerability Score.

6.2 DBSCAN

In the Demographics Cluster Plot (Figure 3), it was seen that smaller towns/districts with low population density were all placed in a single, very large and dense cluster. Also, almost all outliers were larger districts/cities with high population and large area. The Vaccination Cluster Plot (Figure 4) also showed a similar trend. Since the components involved did not include coordinates, the clustering is independent of physical proximity.

7 Conclusions

The final Results obtained are extremely insightful due to their coherence to actual data.
Based on the vulnerability metrics, movement restrictions of corresponding severity can be placed on railway stations as well as state borders.
The various clusters obtained can be used for establishing contact networks between districts in the same cluster. If a counter-measure produces good results in some district, it can also be applied to other districts in the same cluster with a high success probability.
It would be more efficient if Metropolitan cities and larger districts were to come up with independent plans of action. There can be a centralized strategic committee for very small districts and towns which are sparsely populated. Other districts can communicate with their "sister districts" (districts in the same cluster) to come up with effective plans that have shown / could show positive results.

8 References: Datasets

[1] Demographics Dataset: https://livingatlas.esri.in/server/rest/services/LivingAtlas/INDDemography/MapServer/0

[2] Vaccination Dataset: https://dashboard.cowin.gov.in/

[3] Health Infrastructure Dataset: https://pib.gov.in/PressReleasePage.aspx?PRID=1539877

[4] Case Counts: https://github.com/covid19india/api

[5] Indian Railways Network: https://github.com/datameet/railways

9 References

[1] Dave, Raviraj Choudhari, Tushar Bhatia, Udit Maji, Avijit. (2020). A Quantitative Framework for Establishing Low-risk Interdistrict Travel Corridors during COVID-19.

[2] Soni, Shivangi Gauri Shankar, Venkatesh Sandeep, Chaurasia. (2019). Route-The Safe: A Robust Model for Safest Route Prediction Using Crime and Accidental Data. 28. 1415-1428.

[3] Ghosh, Kapil, et al. "Inter-State Transmission Potential and Vulnerability of COVID-19 in India." Progress in Disaster Science, vol. 7, 2020, p. 100114. Crossref, doi:10.1016/j.pdisas.2020.100114.

[4] Flores Tavares, Fernando Betti, Gianni. (2020). Vulnerability, Poverty and COVID-19: Risk Factors and Deprivations in Brazil.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. (pp. 226–231). AAAI Press.