Geo Spatial Health Mapping

Problem Statement

Context

Annual Health Survey (AHS) 2012-13 was conducted in Empowered Action Group (EAG) states i.e. Uttarakhand, Rajasthan, Uttar Pradesh, Bihar, Jharkhand, Odisha, Chhattisgarh & Madhya Pradesh and Assam. These nine states account for about 48% of the total population and 59% of new Births. As per GOI, the purpose of the survey is:

"*The objective of the AHS is to yield a comprehensive, representative and reliable dataset on core vital indicators including composite ones like Infant Mortality Rate, Maternal Mortality Ratio and Total Fertility Rate along with their co-variates (process and outcome indicators) at the district level and map the changes therein on an annual basis. These benchmarks would help in better and holistic understanding and timely monitoring of various determinants on well-being and health of population particularly Reproductive and Child Health*"

GOI also maintains Census Data & Socio Economic Caste Census Data and many of the variables covered in the former overlap with those covered by AHS.

National Health Systems Resource Centre (NHRC) has the data about the state of health facilities in a district. It provides comprehensive information in terms of number of facilities, type/level of facility (PHC/SHC/THC), accessibility, trend (growing/stagnant/depreciating) etc.

Further GOI had categorized districts as-Red, Orange, Green-depending on the basis of severity of Covid-19. Data about spread and effect of Covid-19 in these districts is also available in terms of infected, recovered and deaths.

All these information gathering help policy makers assess the situation on ground, evaluate progress of past/ongoing schemes & efforts and make data driven decisions for the future.

What are the gaps that provide window for improvement?

These data repositories are quite extensive in themselves and provide useful info, but there still are laggings which prevent their utilization to full potential. Following are some of the aforementioned shortcomings/ limitations:

1. Annual Health Survey is limited in coverage. It covers only 9 states. Both Census of India & Socio Economic Caste Census

make available the macroeconomic & demographic data for remaining 20 states and 7 UTs. By deriving the relationship between macroeconomic & demographic data and the composite health indicator, via regression analysis on the AHS data, we would be able to predict the aforementioned indicators for the remaining states.

2. These repositories sit in silos. Most of the analysis done on them is based on their data alone. This could be due to inter departmental compartmentalisations. Using them simultaneously will allow such cross functional analysis which isn't possible using any of the repositories alone. The effect of combining these repositories would not be summation of potential but synergy of potentials.

The question that we wish to answer

AHS Analysis: - As proposed in the point 1 of above segment. Following are the metrics which can predicted at Pan India level.

1. Infant Mortality Rate.
2. Maternal Mortality Ratio.
3. Prevalence of Acute diseases.
    a. Diarrhoea/Dysentery
    b. Acute Respiratory Infection (ARI)
    c. Fever (All Types)
4. Prevalence of chronic diseases
    a. Diabetes
    b. Hypertension
    c. Tuberculosis (TB)
    d. Asthma / Chronic Respiratory Disease
    e. Arthritis

Once predicted, pan India mapping of all these can be done with district level granularity.

Health Infrastructure Analysis: - By using AHS in tandem with NHRC data

1. We can analyse how much the prevalence of a particular disease is dependent on the status & accessibility of healthcare services of the area.
2. Provide actionable insights pertaining to the aspects of healthcare that need to be focussed upon.

3. This can, if time permits, can be scaled to all India level using the output from AHS analysis as input.

Pandemic Response Analysis: - Using the combination AHS, NHRC and Covid-19 Data we'll try to gain insights into how vulnerable an area is to any future pandemic. This analysis would consist of 3 parts.

1. We will explore if there is any relationship between the severity of Covid-19 and the general health profile of the area.
2. How much was the role of health infrastructure in dealing with the pandemic by comparing healthcare infrastructure of two areas with similar health profile but difference in levels of severity (morbidity + mortality) of Covid-19.
3. Above two projects, if time permits, can be scaled to all India level using the output from AHS analysis as input.

Tentative Timeline for the Project

Data Collection, Data Understanding and Data Cleaning

Time Needed: - 1.5 Weeks - (Oct 22 to Nov 2)

We would require data in tabular format. We take AHS data as our frame of reference and would get data from other sources corresponding to columns present in AHS dataset. It should be noted that though we know what columns to look for, yet separate tables need to be explored to find out which of them has our desired feature. AHS, SECC, Census, NHRC and Covid-19 datasets have different column names for the same info and often the column name doesn't give intuition about the info it represents. Hence creation of exhaustive data dictionary is imperative.

1. **Data Collection**
    a. The macroeconomic & demographic data for the states not covered by AHS would have to be downloaded from Census 2011 and Socio Economic Caste Census. Most of the data (Non AHS) needed is present as fragmented tables on their respective websites. By fragmented, I mean, that a given file might contain information about just one feature/column. After download, these files need to be restructured and

collated to create the master dataset that has all the independent features that are present in AHS dataset.

b. Similarly, files containing Covid-19* and NHRC data need to be downloaded, restructured and collated. The only difference is that, this dataset would cover only the states covered by AHS.

2. **Data Cleaning**
   a. There are more than 600 hundred columns in AHS dataset alone. Three columns together constitute info for a district- for ex rural infant mortality rate, urban infant mortality rate and total infant mortality rate constitute the infant mortality rate for the district. So for treating missing values, the information in the other 2 sister columns can be used. Further the data needs to be divided into urban and rural.
   b. For Census & SECC data, ideally the missing value treatment can be done using the mean value of the nearest districts and the quickest way would be to take average at state level. EDA would reveal the incremental benefit of former over the latter and if required data could be passed through the data cleaning pipeline again. More importantly the labelling of the categorical values in the columns need to be harmonized with that of corresponding column in the AHS dataset. Outlier treatment could be done on the basis of intuition and values in other columns for the same district.
   c. Since Covid-19 and NHRC dataset provide info in addition to that of AHS dataset relabelling is not required. The ideal way for missing value treatment here would be dig down further to block/tehsil level data and then aggregate them to district level but it would be quite time consuming as it would require to download and process additional files. Second best would be to regress the given column against a correlated column for ex PHC against SHC & THC to estimate the value for PHC. Easiest way would be to take the state average. Outlier treatment could be done on the basis of intuition and values in other columns for the same district.
3. **Work Assignment**: -Since this is the most time consuming part, whole team will be involved in this process. The process will be partitioned, for ex n team members would be given to work on set of x states. It is yet to be decided, which & how many team members would be assigned to a particular subpart.

Multiple Datasets are present

Exploratory Data Analysis
Time Needed: - 1 Weeks (Nov 3 to Nov 10)

EDA to find out the descriptive statistics and broad trends. Use of geo-pandas is envisaged to gain better insights into data.

1. By this time the workload would have significantly reduced but EDA still would be a time consuming endeavour, as more than 650 columns are involved. Only 2 teams, of three member each, would work on EDA.
2. Team-1 would look into SECC and Census datasets, while Team-2 will work on the NHRC and Covid-19 datasets. Both teams will periodically (every alternate day) share their findings with the entire team and take their inputs.
3. Redundant columns would be identified and removed.
4. Feature creation, for better explaining the variation in dependent columns, would be done. This would include:
    a. Creation of categorical column to classify health profile of a district, for ex: very poor, poor, satisfactory, good, and excellent.
    b. Categorical column to classify the growth of corona cases in a district based of most recent data, for ex steep rise, rising, slow rise, stagnation, steep rise with plateau, rise with plateau, slow rise with plateau, immediate control.
    c. Creation of categorical column to classify state of health infrastructure in a district, for ex: very poor, poor, satisfactory, good, and excellent.

Dimensionality Reduction and Clustering
Time Needed: - 4 Days (Nov 11 to Nov 14)

Owing to large number of columns dimensionality reduction is imperative to obtain meaningful clusters. Clustering would be done to identify similar districts.

1. Above steps would be done for each of the analysis mentioned in the problem statement
2. Selection of dimensionality reduction algorithm.
3. Selection of clustering algorithm and benchmarking of clusters.
4. This step can be handled by 1 team alone and if needed 2 teams.

Model Creation and Selection

Time Needed: - 3 Days (Nov 15 to Nov 18)

Creation of a multi label regression model which can predict the complete health profile of the given population, as per its characteristics. Use of Multi Output Regression library of sklearn is envisaged. For the second use-case in the pandemic response analysis descriptive regression analysis would be performed and preference would be given to simpler algorithms

1. For each analysis, model creation would entail creation separate model for each cluster.
2. Collation of prediction findings into a single file.
3. This step can be handled by 1 team alone and if needed 2 teams.

Integration with Power BI

Time Needed: - 3 Days (Nov 19 to Nov 22)

The finding with actionable insights would be presented via interactive visualization using Power BI.

1. Different dashboards/tabs would represent each analysis.
2. Python integration with Power BI is envisaged, to automate the whole process, is envisaged.