National Rural Drinking Water Programme (NRDWP)

ARUN INANI
2017A7PS0085H
Department Of Computer Sc.
BITS Pilani Hyderabad Campus
f20170085@hyderabad.bits-pilani.ac.in

ABHISHEK BHARDWAJ
2017A7PS1497H
Department Of Computer Sc.
BITS Pilani Hyderabad Campus
f20171497@hyderabad.bits-pilani.ac.in

AKSHAT GUPTA
2017A7PS1699H
Department Of Computer Sc.
BITS Pilani Hyderabad Campus
f20171699@hyderabad.bits-pilani.ac.in

I. PROBLEM MOTIVATION

We aim to determine that the potable water availability across India depends on the location or category wise population distribution.

This analysis will help us to find solid proofs (if any) to the common myth that water availability depends upon population distribution in that area and is much better for general category people compared to other classes.

II. BACKGROUND

The National Drinking Water Programme (NRDWP), is the Government of India's (GoI's) flagship rural drinking water supply scheme. The scheme aims to provide safe and adequate water for drinking, cooking, and other domestic needs on a sustainable basis.

The Central Government assistance to States for rural water supply began in 1972 with the launch of the Accelerated Rural Water Supply Programme. It was renamed as National Rural Drinking Water Programme (NRDWP) in 2009, which is a centrally sponsored scheme with fund sharing between the Centre and the States. Under NRDWP, one of the objectives was to "enable all households to have access to and use safe & adequate drinking water within premises to the extent possible". It was proposed to achieve the goal by 2030, coinciding with the United Nation's Sustainable Development Goals.

Dataset link: *Basic Habitation Dataset*

III. OBJECTIVES

- To find the availability of potable drinking water (State-wise distribution)
- To find the availability of potable drinking water (Category wise distribution)
- To check whether the availability of water is dependent upon the location.
- To check if there is any difference in availability of water depending upon category.
- To compare the effect of location and category on water availability.
- To compare performance of different states of India.

IV. METHODOLOGY

A. Data Preprocessing

Below listed Data preprocessing techniques were used to ensure best result and better analysis.

1) *Data Cleaning:*
   a) *Removal of any duplicate data entries.*
   b) *Noise Removal:* Removal of rows having any of it's attribute entry less than zero or all attribute entries equal to zero. Also removed rows having covered population greater than current population for any of it's three categories.
2) *Dimensionality Reduction:*
   Removal of unnecessary or unused columns viz. Year, Status, SC Concentrated, ST Concentrated.
3) *Feature Construction:*
   Created 4 new features for better data representation viz. SC covered population/SC Current Population And similarly, for ST and General category and Total Fraction (Total Covered Population/Total Current Population).
4) *Aggregation:*
   Aggregation was done to obtain data at district level and to use the newly created four features. Similarly, aggregation was done at state level and national level.

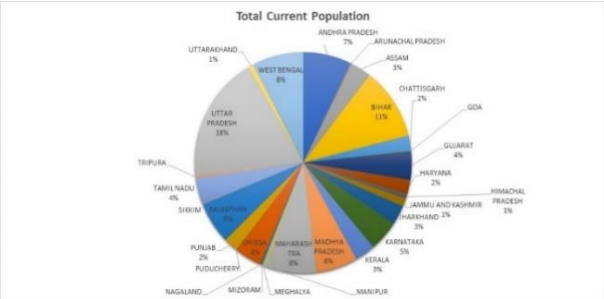
B. Mining of Data

1) *Multilevel Clustering:*
   Clustering algorithm was aplied on various level viz, National,State and District to search for any category wise or locationwise distributionn.

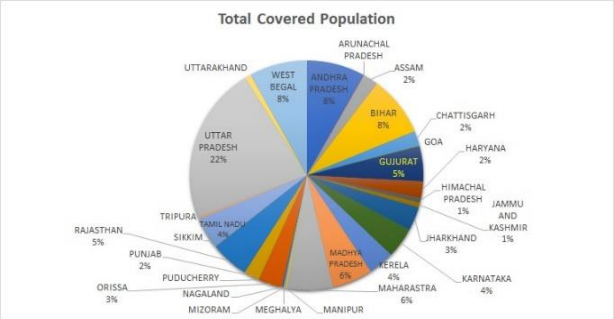
2) *Outlier Analysis:*
   Outliers found at different levels were studied throughly to find the cause of their unusual behavior.

V. DATA VISUALIZATION:

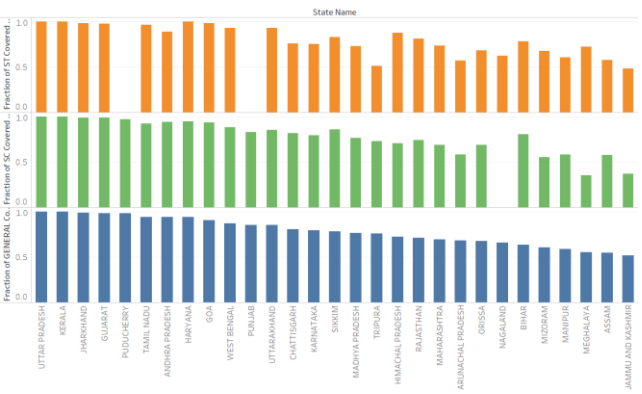
The below-shown pie chart shows the percentage of total current (total population of India) population of India residing in a particular state for all the states and UTs.



As it is quite evident from the chart that almost 45% of population resides in 5 states i.e. UP, Bihar, WB, Maharashtra, and MP. And the remaining 55% population has major contribution from Rajasthan, Orissa, AP, and less than 1% contribution from states like Tripura, Assam, and Nagaland.



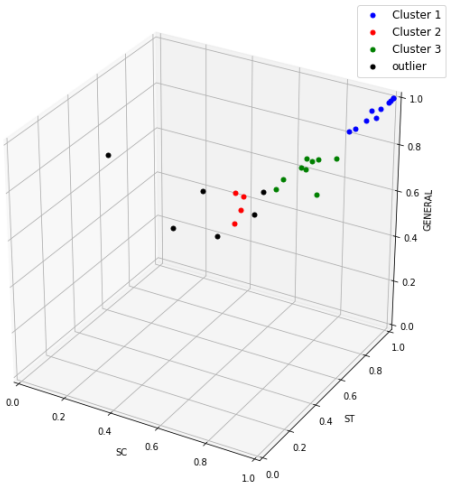
The above-shown pie chart shows the percentage of total covered (population having access to potable drinking water) population of India residing in a particular state for all the states and UTs. As it is visible from this almost 45% of the total population having access to potable drinking water resides in UP, WB, Bihar, Maharashtra, and MP (same states which had a major share in total current population). Also, there are certain states like AP, Gujarat, and Kerala which have a higher percentage of covered population compared to the current population indicating that these states might have a higher fraction of people having access to potable drinking water. Also, then there are states like Bihar and Orissa which had a higher percentage of the total population but have a lower percentage in people having access to water hence might result in a lower fraction of people having access to potable drinking water.



This Bar graph shows the fraction of people having access to potable drinking water for all states Category-wise. The top row (orange) shows the fraction of ST Category, Middle row(green) shows the fraction of SC Category, Last row (Blue) shows the fraction of General Category. As can be seen, there is a group of sates with a high fraction (greater than 0.9, almost equal to 1) of all the three categories like UP, Kerala, etc. Then there is a group of states like Rajasthan, Maharashtra with values of all three fractions in range (0.55-.75). There is one more group of states like Assam and Manipur having all the fractions in range (0.45-.55) with no significant difference among the three categories. Then there are certain states like Punjab, Nagaland, J&K having no similarity with any other state/UT.

For better visualization, a scatter plot was constructed with X-Axis as the fraction of SC category having access to potable drinking water, Y-Axis as the fraction of General category having access to potable drinking water and, Z-Axis as the fraction of ST category having access to potable drinking water (*This axis convention will be followed throughout the report*).

NATIONAL LEVEL



DBSCAN was applied to this plot to find group of states with similar fractions, this gave us 3 clusters and a bunch of outliers:

- Cluster1(Blue): This cluster had states like Goa, Gujarat, Haryana, Jharkhand, Kerala, TN, UP, UK, WB. All this state had all three fractions in the range (.8-1.0) depicting these states performed better than all other states in terms of availability of water and here availability is not dependent on a category (For state level aggregation).
- Cluster2(Red): This cluster had states like Arunachal Pradesh, Assam, Manipur, and Mizoram. All these states have values of three fractions in range (0.4-0.6),

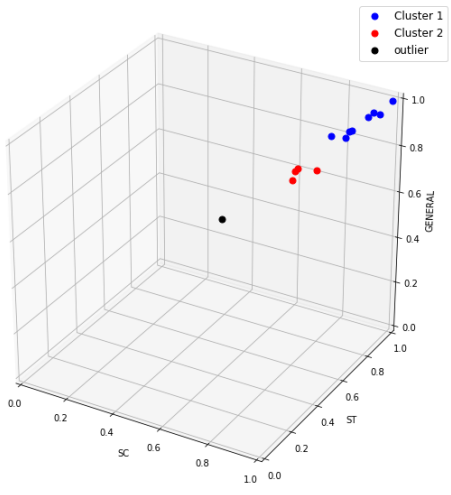
with values of the fraction of ST covered population being slightly higher than the other two categories fraction covered.Cluster2 comprising of 4 North-Eastern states suggests that the availability of potable drinking water is slightly dependent on location and with the location factor having a negative effect on North-Eastern states.

- Cluster3(Green): This cluster comprises of states like Bihar, Chhattisgarh, HP, Karnataka, MP, Maharashtra, Orissa, Sikkim and Rajasthan. These states have values of all the three fraction in the range(0.6-0.8) and have no significant difference in the values of fraction of all three categories, implying the fact that availability of water in these three states do not depend on a category for these states(on state level aggregation), also since all these states belong to different demographic parts of country hence this cluster is not location dependent.
- Also, there are certain outliers like J&K, Meghalaya, Nagaland, Puducherry, Punjab, and Tripura. These states are not a part cluster because most of these states have values of fraction of one of the categories equal to zero or have the value of all the three fractions much lower compared to any other cluster showing very poor condition of availability of water in these states.

CLUSTER 1

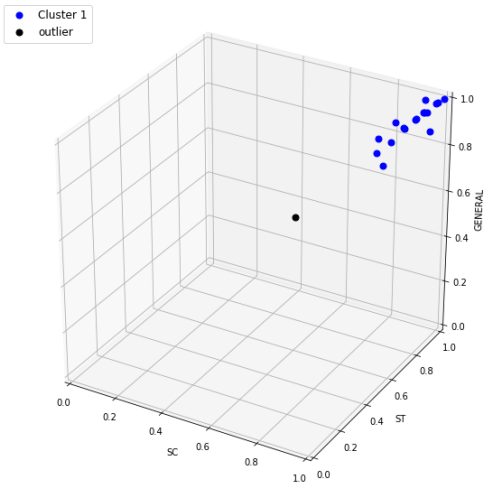
As seen earlier cluster1 was a group of states having high fraction values for all the three categories indicating no appreciable category wise distribution (on state level after aggregation). To get a better insight into distribution (if any) we studied two states of cluster1 Uttarakhand and West Bengal.

UTTARAKHAND



The above graph shows the scatter plot of the various districts in Uttarakhand with axis convention as defined earlier. DBSCAN was applied to this plot and it resulted in 2 clusters and one outlier. For every district in cluster1 values of all three fractions lie in the range (0.75-0.95) whereas values of fractions for the district in cluter2 lies in the range (0.55-0.7). Even though DBSCAN resulted in two clusters, all the districts don't differ from each other significantly owing to the high fraction value (district level) for all the three categories. Hence, Uttarakhand shows no category of wise destruction.

WEST BENGAL



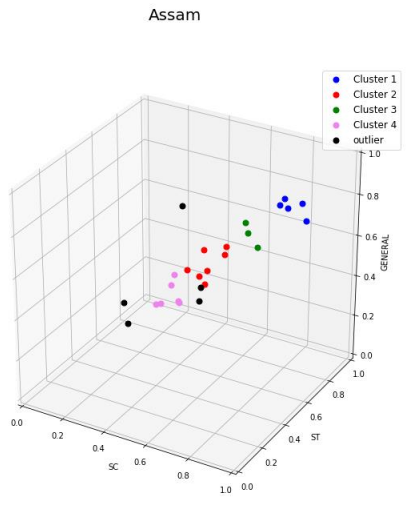
The above graph shows the scatter plot of various districts in West Bengal with axis convention defined earlier. DBSCAN was applied to this plot and it resulted in 1 cluster and an outlier point. As seen earlier, West Bengal is a part of Cluster1(national level) hence all the districts of West Bengal have high values for all the three fractions in the range (0.75-1.0). There is no significant difference between the values of all the districts hence there is no possibility of any sort of category-wise or location-wise distribution.

All the other states of Cluster1 viz. Goa, UP, Kerala, etc. follow the same path with no appreciable difference in district values hence indicate towards the fact that cluster1 states have no significant category wise distribution nor any location-wise distribution. But in general, these states have better availability of potable drinking water compared to other states/UTs in India.

CLUSTER 2

As seen earlier cluster2 was a group of states having a low fraction of total covered population in the state (Total covered population/Total current population) with no appreciable category wise distribution (on state level after aggregation). But this cluster does have high inter-cluster similarity between all four states viz. Arunachal Pradesh, Assam, Manipur, and Mizoram as all of them have a very low percentage of people having access to potable drinking water (State level).

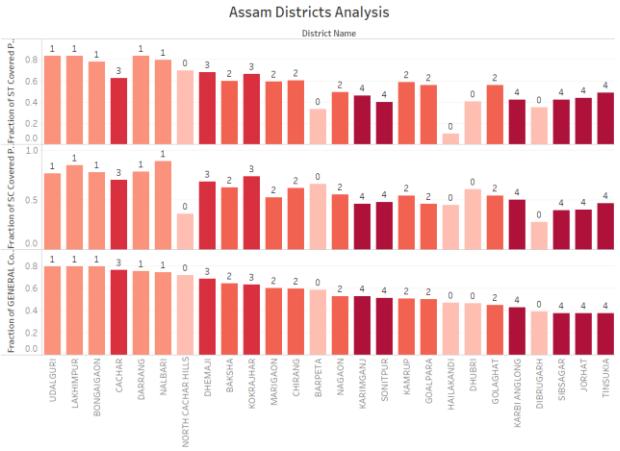
To get a better insight into distribution (if any) we studied Assam's distribution on the state level. The below graph shows the scatter plot of various districts in Assam with axis convention defined earlier.



DBSCAN was applied to this plot and it resulted in 4 clusters and 5 outliers.

- Cluster1(Blue): It comprises all the districts that have a higher percentage of people having access to water compared to other remaining districts. All three fractions of these districts lie in the range (0.70-0.85), with no significant difference among the three values implying the fact that districts of cluster1 have no Category-wise Distribution.
- Cluster2(Red): This cluster consists of those districts having values of all three fractions in the range (0.55-0.65). All three categories of these districts have almost equal access to potable drinking water.
- Cluster3(Green): It comprises of all the districts having values of fraction in the range (0.65-0.80) with no significant dominance of any category above remaining categories and hence no category-wise distribution.
- Cluster4(Pink): This cluster has all those districts with the poorest values of fraction of people having access to potable water, values in this cluster lie in the range (0.45-0.55).
- Outliers (Black): There were certain districts like Barpeta and Dhubri where one of the categories had a better fraction than the remaining two values of all these districts vary largely. Districts of outliers have a moderate category-wise distribution with ST performing the poorest followed by SC and then general.

So, it can be safely concluded that the general category has better access to potable drinking water than SC, ST in all outlier districts.

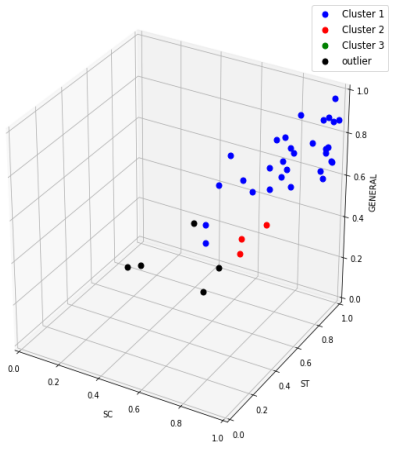


The above Bar Graph shows the fraction of SC population (Total SC covered/Total SC current) and similarly ST population and General population for each District. Bar graphs with label 0 denote outlier districts and with label 1,2,3,4 denotes cluster 1,2,3,4 respectively. This Bar graph shows that districts of cluster 1 have the highest access to potable drinking water, followed by cluster 3, cluster 2, and cluster 4. As it can be seen there is no significant difference between the three fractions for the districts, it can be concluded that there is no category-wise distribution.
Also, these clusters differ mostly due to location factors owing to other sub-factors like their geographical location, average rainfall, etc. Since all these districts as shown earlier have no significant category-wise distribution (due to almost similar values of fraction of all three categories in most districts), the distribution on state level can be only because of the location factor.
Now as Arunachal Pradesh, Assam, Manipur, and Mizoram all were part of cluster2(National level), they all follow the same trend as that of Assam and have very little category wise distribution on the district level.

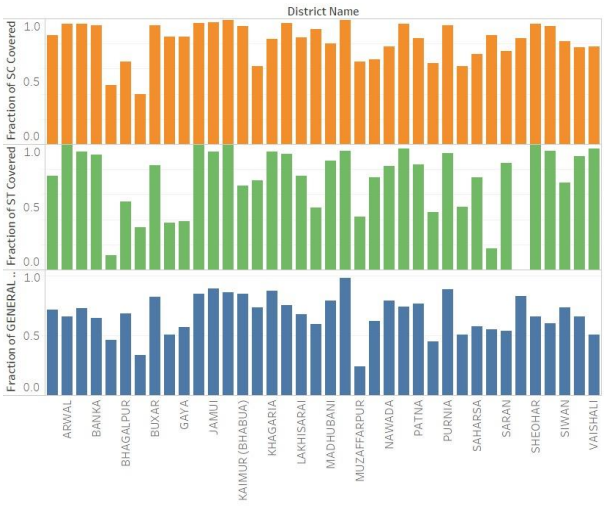
CLUSTER 3

It consists of states with the value of total fraction (Total current population/Total covered population) higher than cluster2 but lower than cluster1, with values is in range (0.65-0.80). These states account cumulatively for almost 40% of India's population and comprise 36% of India's Population having Access to water, this cluster is the second-best performing cluster on the national level. These states have very little difference in values of fraction of all three categories (State level average). Thus, have no conclusive proof of location-wise or Category-wise distribution.
For better insight, Bihar was analyzed on the district level. The below-shown graph is a scatter plot of all the districts in Bihar with the same axis convention used before. DBSCAN was applied to this plot, which resulted in two clusters and 5 outliers. Cluster1(Blue) is a set of districts that performed better than other districts in the state and have values of total fraction in the range (0.6-0.9) and cluster2 comprises of districts with values in the range (0.4-0.6). These two clusters differ mainly due to difference in fraction of covered population. (district level).

BIHAR



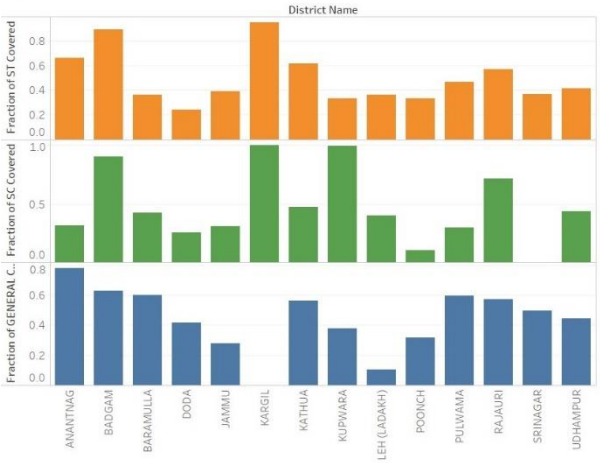
Outliers observed existed mainly due to the reason that one category fraction has a very low value compared to other categories, and since all categories(i.e. SC, ST, General) has the lowest fraction in any one of the outlier district there is no proof of any category wise distribution in outliers also.



Above shown is a bar graph of all the three fractions for all the districts of Bihar. As seen from the cluster plot here also there are two types of districts (cluster1 and cluster2) and some outliers. Certain districts performed well and some districts with values closer to that of state level values. Thus, it can be safely concluded that there is no significant category/location wise distribution in Bihar at the state level.
Further analysis of different states of cluster3 like Rajasthan, MP, Orissa, etc. resulted in the same type of conclusions with no appreciable category-wise distribution among states on the district level. Below attached are Bar graphs of Maharashtra and Orissa showing values of fraction of all three categories in different districts of Maharashtra and Orissa respectively.

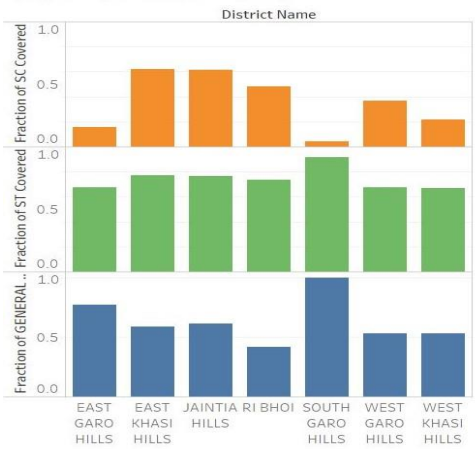
OUTLIERS

As seen earlier there were 6 outliers after DBSCAN was applied on the National level namely, J&K, Meghalaya, Nagaland, Puducherry, Punjab, Tripura. Some of them were due to the fact, that one category being heavily dominated by the other two categories in terms of the fraction of people having access to potable water in that particular category, examples of this are Meghalaya, Nagaland, Puducherry, Punjab.
J&K and Tripura were outliers because they had the lowest total covered population fraction (Total population covered/Total current population) in the country.



Above shown is a Bar graph of three fractions of various districts of J&K. State level average value of fraction of SC (0.36), ST (0.48), General (0.51) indicate a very light effect of category wise distribution in state.
Values of the three fractions for different districts differ greatly. Most of the districts like Anantnag, Pulwama, Baramula, etc. have a significant difference in the values of their respective categorical fraction, with the value fraction of general being the highest one among all and Sc being the lowest. Some districts like Badgam, Kathua, etc. have almost similar values of their respective categorical values. Then there are certain unusual performing districts like Kargil and Srinagar with value of one categorical fraction equal to zero. Some districts do have highest fraction of SC category.
But overall availability of potable drinking water in J&K is dependent on category-wise distribution.

Below shown is a Bar graph of three fractions of various districts of Meghalaya. On state level Fraction of ST (0.72) and Fraction of general (0.55) having access to potable drinking water was almost double that of SC (0.35). But when we had a closer look at different districts 5(South Garo Hills, East Garo Hills, West Khasi Hills) out of 7 had particularly high values of fraction of ST, followed by the fraction of General and then by Fraction of SC. The remaining two districts had almost similar values for all the fractions. Hence it can be safely concluded that there is a category-wise distribution in Meghalaya with the fraction of SC being lowest.

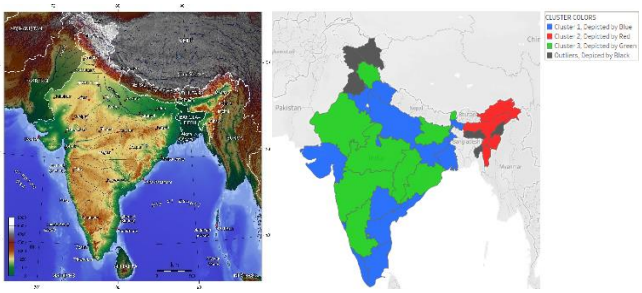


Also, states like Punjab and Puducherry perform very well on the national level in terms of total fraction, but internally have zero fraction of ST category, hence they were not a part of any cluster on the national level. This is mostly because of the incompleteness of dataset and hence no conclusions can be drawn for these states.

VI. CONCLUSION

We analyzed the availability of potable drinking water in India on many levels to find any sort of category-wise or location-wise distribution. Starting from the National Level we calculated values of the three fractions for each state and made a 3D scatter plot of the same which resulted in 3 clusters. They had no category-wise distribution but had a slight location-wise distribution which resulted in cluster2, which comprises of Arunachal Pradesh, Assam, Manipur, and Mizoram, all of which have lowest total fraction values in India. Then to have a better insight we studied each cluster on the state level which led us to observations:

- Cluster1 have no location-wise or category-wise distribution, but every state in cluster1 has value of Total fraction significantly higher than states which are part of other clusters.
- Cluster2 states have some of the lowest total fraction (state level) values. On analyzing them on state level we found little bit traces of category-wise distribution in couple of districts and a weak location-wise distribution on state level.
- Cluster3 states have fairly high values of total fraction, but are lower than that of cluster1 state values. States of cluster3 have no significant category-wise or location-wise distribution on any level. At all possible levels fraction of all categories don't differ by a lot.
- Outliers found after National level clustering were analyzed deeper and states like J&K which had the lowest total fraction in the country has a little category-wise distribution. Whereas states like Punjab, Puducherry, etc. have shown particularly different behavior as compared to other states with the fraction of one category having value zero on every possible level (state, district, village) which is mostly because of data inconsistency/incompleteness. Whereas Meghalaya had a strong category wise distribution on the state and district level.



So, after all this analysis it is quite evident that there is a significant location-wise distribution in the availability of potable drinking water in India due to its large land size and varying demographic and geographic conditions. Above shown is a Topological map of India, on comparison with the cluster map shown just above, it is visible that the cluster formed do coincide with the topological map.Cluster1 states are a part of Indo-Gangetic Plains and some part of the East coast, whereas Cluster3 states are a part of the central Highlands and West Coast, Cluster2 states are completely separated from others in the Northeast Mountain ranges. All these points do support our facts of location-wise Distribution in India on National level.
Also, there is no sort of Category-wise distribution present in India for the availability of potable drinking water except a couple of cases which rules out any sort of category -wise discrimination on category basis for water availability.

VII. SOFTWARES USED

1) Jupyter
2) Tableau

****For GitHub repo click here****