EasyChair Preprint

№ 5350

Statistical Analysis of Semi-Urban Districts of India

Yash Dalal, Arindam Chaudhuri and Jaykrishna Joshi

April 18, 2021

STATISTICAL ANALYSIS OF SEMI-URBAN DISTRICTS OF INDIA

Yash Dalal
Mukesh Patel School of Technology Management and Engineering
NMIMS University Mumbai
yash.dalal1721@nmims.edu.in

Arindam Chaudhuri
Mukesh Patel School of Technology Management and Engineering
NMIMS University Mumbai
arindam.chaudhuri@nmims.edu

Jaykrishna Joshi
Mukesh Patel School of Technology Management and Engineering
NMIMS University Mumbai
jaykrishna.joshi@nmims.edu

Abstract: The semi-urban districts hold a lot of importance in the economy of India. They have an economy consisting of primary, secondary as well as tertiary sectors. Rural areas on the other hand have a primary sector based economy and urban areas are focused upon tertiary sector. The objective of this project is to find out the factors that can lead to the development of these semi-urban districts of India. This analysis will also focus upon the things which can be improved for further enhancing the growth in these regions.

Keywords: semi-urban, hierarchical clustering, data analysis

1 Introduction

In India, there are total 718 districts in 28 states and 8 union territories. These 718 districts can be broadly classified into three types – urban, semi-urban and rural. According to Collins dictionary, semi-urban is an area which is not fully urban nor fully rural. It has a mixture of rural and urban characteristics [1]. Semi-urban districts are those which have an urban population between 35% - 60%. Semi – urban districts have an economy which is based on primary, secondary as well as tertiary sectors. This gives it a rural as well as an urban aspect. The economies of urban areas are mostly confined to tertiary sector while rural areas have more focus upon primary sector. There a total 75 semi-urban districts in India.

The main aim of this project is to find relationships between the variables and to know the extent of influence on each other. Most of the insights of this project is based upon unsupervised machine learning techniques. When the data points are segregated into clusters it becomes easy to find factors which can be improved upon.

Detailed analysis of semi-urban districts of India' is a data analysis and clustering project. The most important task of this project is data mining and data pre-processing. It is often said that 80% of the time in any data analytics project is consumed by data mining and remaining for data pre-processing, exploratory data analysis and algorithm training. Extensive Exploratory Data Analysis is performed to get an in-depth knowledge of the data. Data Visualization which is rightly called Data Storytelling reveals a lot of information in itself. Moreover, hierarchical clustering method is used widely in this project to identify the cluster of districts facing similar kind of issues. The results from the clustering methods will help to identify the key problems faced by the districts in the cluster.

This paper is organised as follows. In section 2 related work is presented. This is followed by data mining activities in section 3. The mathematical model is given in section 4. In section 5 methodology is highlighted. The results are placed in section 6. In section 7 conclusion is given.

2 Related Work

There has not been much research into the economical, health, demographic and developmental aspects of semi-urban districts in India. In the review paper [2], the authors have focused upon the semi-urban areas also called as peri-urban. Their focus is to discuss the ways in which these areas can be developed so that it will have lesser environmental impact and can be called as sustainable development. The authors have mentioned many research programs which are undertaken by nations to find out an optimal solution to develop the semi – urban areas to urban areas in future without much damage to the environment. In this paper, many developmental models used to develop the urban fringe or semi – urban areas have been discussed.

The research paper [18], talks about the challenges faced by the peri – urban areas surrounding the megacities in India. Peri – urban area is the acronym for the area which is on the periphery of an urban area. The paper discusses two main points – problems and the changes occurring in peri – urban areas of India and the approach taken by the government agencies towards planning and managing these areas. The problems identified by the author are – unregulated development, population displacement, land use conversion, poor sanitation and lastly poor mobility and connectivity.

Author in the paper [19], discusses the environment, social, livelihood, water, technology and ecology aspects of the peri – urban areas. The main task in this discussion paper is to impart information on the above aspects with respect to peri – urban areas. All the points are explained in detail by the author. This paper is a part of discussion series.

The research paper [20] maps the rural, urban and semi – urban areas according to the density grids. The task undertaken by the author is to classify European Union (EU) into the above-mentioned areas. The population and area of the commune (the definitions differ from country to country) have been taken into consideration to find out the density of population. On the basis of density of population, the author segregated the areas in Europe into rural, urban, semi – urban and remote. However, the author asserts that there is lack of proper threshold to segregate such areas.

3 Data Mining Activities

The data for this project is collected from various sources such as Central and State government websites, third party websites, district websites, websites of the ministries and newspaper articles. The data has been segregated into demographic, health, economic, water quality and land use.

The chart on below in Fig. 1 shows the representation of semi - urban districts from each state. The more industrialized and high population states have larger proportion of semi - urban districts. 11 districts of Tamil Nadu are categorized as semi - urban. It is followed by Gujarat, Uttar Pradesh, Maharashtra and Karnataka at 6 each. Another important factor to be considered is that the area of the districts in Tamil Nadu is comparatively smaller than that of the other states. In total 75 districts have been considered for this project; they are as follows:

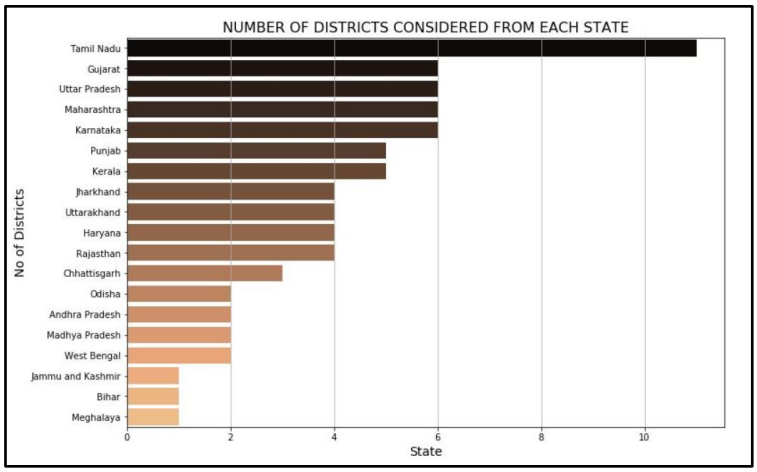


Fig 1. No of districts considered from each state

State	No of Districts	Districts considered
Kerala	5	Kasaragod, Malappuram, Alappuzha, Kollam, Thiruvananthapuram.
Tamil Nadu	11	Karur, Theni, Madurai, Tirunelveli, Salem, Tiruchirappalli, Thoothukkudi, Erode, Namakkal, Virudhunagar, Vellore.
Karnataka	6	Dharwad, Bellary, Shimoga, Mangalore, Mysore, Gadag.
Andhra Pradesh	2	Visakhapatnam, Krishna (Vijayawada).
Maharashtra	6	Akola, Amravati, Chandrapur, Aurangabad, Nashik, Raigad.
Gujarat	6	Gandhinagar, Rajkot, Jamnagar, Bhavnagar, Valsad, Vadodara.
Madhya Pradesh	2	Ujjain, Jabalpur.
Chhattisgarh	3	Durg, Korba, Raipur.
Odisha	2	Sundergarh, Khorda.
West Bengal	2	Darjeeling, Hooghly.
Jharkhand	4	Bokaro, Dhanbad, Ranchi, Jamshedpur.
Bihar	1	Patna
Uttar Pradesh	6	Bareilly, Agra, Varanasi, Meerut, Jhansi, Gautam Buddha Nagar
Haryana	4	Panipat, Ambala, Rohtak, Yamunanagar.
Rajasthan	4	Jaipur, Ajmer, Jodhpur, Kota.
Punjab	5	Ludhiana, SAS Nagar (Mohali), Patiala, Jalandhar, Amritsar.
Uttarakhand	4	Dehradun, Haridwar, Udham Singh Nagar, Nainital.
Jammu and Kashmir	1	Jammu
Meghalaya	1	East Khasi Hills (Shillong)
19 States	75	Total

It is to be noted that there are 2 districts of Telangana (RangaReddy and SangaReddy) that fall under this category, but as they were carved out from larger districts when Telangana was formed in 2014, historical data cannot be accurately extracted. Bardhaman (Burdwan) district of West Bengal was separated into 2 new districts in 2017, therefore data for the same cannot be extracted accurately [4], [5], [6], [7].

Demographics data – [3]

Sr No	Main columns	Sr No	Derived Columns
1	District	1	Percentage of urban population (in %)
2	State	2	Total population growth (in %)
3	Area (in sq.km.)	3	Male population growth (in %)
4	Total population 2011	4	Female population growth (in %)
5	Male population 2011	5	Total literacy growth (in %)
6	Female population 2011	6	Male literacy growth (in %)
7	Urban population 2011	7	Female literacy growth (in %)
8	Rural population 2011	8	Child population proportion (in %)
9	Total literacy 2011	9	Sex Ratio growth (in %)
10	Male literacy 2011		
11	Female literacy 2011		
12	Total population 2001		
13	Male population 2001		
14	Female population 2001		
15	Total literacy 2001		
16	Male literacy 2001		
17	Female literacy 2001		
18	Sex Ratio 2011		
19	Sex Ratio 2001		
20	Density of population (per sq. km.)		
21	Working class population (in %)		
22	Child population under 6 years 2011		

Health data – [8]

Sr No	Main Columns	Sr No	Derived Columns
1	Infant Mortality Rate 2011 (per thousand)	1	Growth in Infant Mortality Rate (in %)
2	Infant Mortality Rate 2001 (per thousand)	2	Growth in Under 5 Mortality Rate (in %)
3	Under 5 Mortality Rate 2011 (per thousand)	3	Population per Government Hospital
4	Under 5 Mortality Rate 2001 (per thousand)		
5	Number of Government Hospitals		

Economic data – [9]

Sr No	Main Columns
1	Below poverty line population (in %)
2	Per capita income (in Rs)

Land Use data – [10]

Sr No	Main Columns	Sr No	Derived Columns
1	Net sown area (sq.km.)	1	Percentage of area sown (in %)
2	Forest area (sq.km.)	2	Percentage of forest land (in %)
3	Net irrigated area (sq.km.)	3	Percentage of sown area irrigated (in %)
4	Rainfall (in mm.)		

Ground Water Quality data – [11]

Sr No	Main Columns
1	Salinity (> 3000 µS/cm)
2	Chloride (> 1000 mg/litre)
3	Fluoride (> 1.5 mg/litre)
4	Iron (> 1 mg/litre)
5	Arsenic (> 0.05 mg/litre)
6	Nitrate (> 45 mg/litre)
7	Total ground water quality

4 Mathematical Model

The formulae used to compute the derived columns are given as below:

1. Urban population (%) = $\frac{(urban\ population * 100)}{total\ population\ 2011}$

   This gives the percentage of population in that particular district which live in urban areas.

2. Population growth (%) = $\frac{(total\ population\ 2011 - total\ population\ 2001)}{total\ population\ 2011} * 100$

   The population growth from 2001 to 2011 is given by this formula. The percentage rise of total population seen by a district can be calculated.

3. Male Population growth (%) = $\frac{(male\ population\ 2011 - male\ population\ 2001)}{male\ population\ 2011} * 100$

   Similar to the total population growth formula, this only takes into account the rise in male population between 2001 and 2011.

4. Female Population growth (%) = $\frac{(female\ population\ 2011 - female\ population\ 2001)}{female\ population\ 2011} * 100$

   The formula is the same as Male population growth, just the difference here is that it calculates the growth in number of females during a given time.

5. Literacy growth (%) = $\frac{(total\ literacy\ 2011 - total\ literacy\ 2001)}{total\ literacy\ 2011} * 100$

The growth in literacy rate from 2001 to 2011 is given by this formula. The growth is calculated in terms of percentage.

6. Male Literacy growth (%) = $\frac{(male\ literacy\ 2011 - male\ literacy\ 2001)}{male\ literacy\ 2011} * 100$

The growth in male literacy rate is computed using this formula.

7. Female Literacy growth (%) = $\frac{(female\ literacy\ 2011 - female\ literacy\ 2001)}{female\ literacy\ 2011} * 100$

Similar to the previous two formulae (5 and 6), this also calculates growth rate but for female literacy levels between 2001 and 2011.

8. Child population proportion (%) = $\frac{Child\ population * 100}{total\ population}$

The above formula gives the proportion of children below 6 years of age with respect to the total population.

9. Sex ratio growth (%) = $\frac{(sex\ ratio\ 2011 - sex\ ratio\ 2001)}{sex\ ratio\ 2011} * 100$

Sex Ratio Growth formula is similar to population and literacy growth formulae. It calculates the percentage growth in sex ratio levels between the given time frame.

10. Growth in IMR (%) = $\frac{(IMR\ 2011 - IMR\ 2001)}{IMR\ 2011} * 100$

IMR is the acronym for Infant mortality rate. The above formula gives the percentage increase or decrease in the Infant Mortality rate between a given time frame.

11. Growth in U5MR (%) = $\frac{(U5MR\ 2011 - U5MR\ 2001)}{U5MR\ 2011} * 100$

U5MR is the acronym for Under 5 Mortality Rate. The formula similar to IMR, computes the increase or decrease in U5MR levels.

12. Population per Government hospital = $\frac{Total\ population\ 2011}{Number\ of\ Government\ Hospital}$

Population per Government Hospital gives the number of people covered by one single government hospital.

13. Percentage of area sown (%) = $\frac{(Net\ sown\ area * 100)}{Total\ area}$

The percentage of area utilized for agricultural purposes out of the total land area available in that district.

14. Percentage of area under forests (%) = $\frac{(Forest\ area * 100)}{Total\ area}$

The percentage of area under forest out of total area of that district.

15. Percentage of sown area under irrigation (%) = $\frac{(Net\ irrigated\ area * 100)}{Net\ sown\ area}$

Proportion of area under irrigation out of total area under agriculture.

The project mostly concentrates on exploratory data analysis (EDA), data visualization and clustering algorithms. The following paragraphs will throw a light on the basic understanding of these concepts.

Exploratory Data Analysis (EDA): The definition of Exploratory Data Analysis states that it is the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. EDA helps to gather information about the data before actually applying algorithms on it [12].

Data Visualization: In simple terms, data visualization is graphical representation of data. These visualizations are useful to find insights, trends, outliers, etc. There are various types of visualization graphs in which data can be represented, such as scatter plots, bar plots, box plots, histograms, so on and so forth. They usually reveal much more information than numbers can [13].

Clustering: Clustering is a method to find similar groups in a dataset. There are many types of clustering algorithms. First are the connectivity models which consists of hierarchical clustering, where one single cluster is further partitioned into many based on the distances between them. Usually, these models are very easy for interpretation. Second are the centroid models, they are based upon the distance of that particular point from the centroid of the cluster. K-means is an algorithm using this technique.

Hierarchical clustering: Hierarchical clustering starts with the formation of one distinct cluster for one individual point. Then these points are merged into same clusters after computing the distance between them. The decision on the number of clusters to be chosen can be inferred from a dendrogram. The best choice of the number of clusters is based upon the horizontal line that traverses maximum distance without being further divided. The biggest advantage of hierarchical based clustering algorithm is that the results are reproducible (they do not change even after multiple runs through the code) [14].

Metrics: The two commonly used metrics for distance measurement are Euclidean and Manhattan distances. The formulae for both of them are as given below.

1. Euclidean distance = $(\Sigma (a_i - b_i))^{½}$

2. Manhattan distance = $\Sigma |a_i - b_i|$

Apart from these two, there are also some more metrics used in computing distance such as the Mahalanobis distance, Squared Euclidean distance and Maximum distance [14].

Agglomerative clustering methods: Agglomerative clustering method is a part of hierarchical clustering. In this method, all the data points are considered as separate clusters. These are then combined based on the distance between them until they merge to form a one single cluster. This can be represented in a diagram known as dendrogram as mentioned earlier. Agglomerative clustering can further be divided into four distinct types [15]:

Complete linkage clustering – In this technique all pairwise dissimilarities are computed and then the largest values between the clusters is considered as the threshold for dividing the data points. Complete linkage method tends to produce more compact clusters [16].

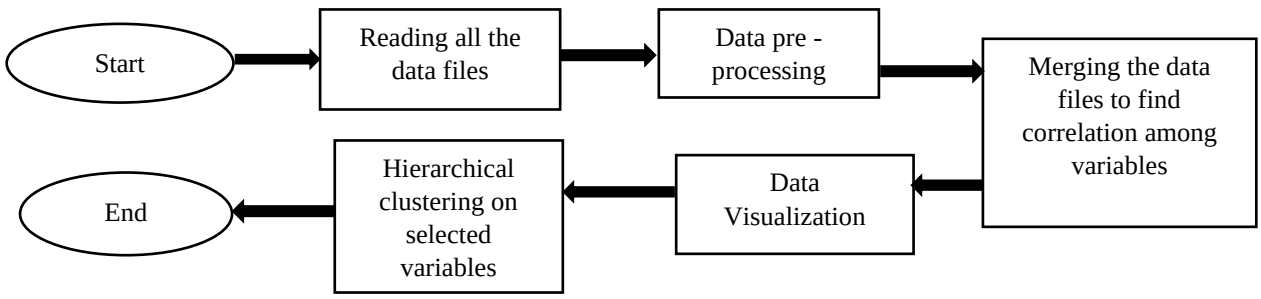
Single linkage clustering – In this type of agglomerative clustering technique, the pairwise distances for the two closest points is computed and then it is combined to form one cluster. It usually creates long and loose clusters. The drawback of this method is that points which are very farther from each other may also be considered in the same cluster [17].

Average linkage clustering – Average linkage clustering calculates all the pairwise dissimilarities between the points and then takes average of them.

Ward's minimum clustering – This method minimizes the within cluster variance. At each step the clusters with minimum between-cluster distances are merged [15].

5 Methodology

Algorithmic flowchart conveys the steps involved in making of a project. This project consists of five main steps and they are as follows:



Step 1: The data gathered has been segregated in five different Excel (comma separated value) files. The files are-

a. Demographics data
b. Economic data
c. Health data
d. Ground Water Quality data
e. Land Use data

Step 2: Data pre - processing forms an important step in any data science project. This step includes renaming the columns, finding out the null values if any, identifying the unique column from all the five files.

Step 3: On the basis of the unique primary key column (the name of the district in this project as it is unique), the csv files are merged. Computing the summary statistics, correlation heatmaps and extensive exploratory data analysis are the tasks performed. EDA helps to understand the data and what relationships can be established.

Step 4: Data Visualization is also called Data Storytelling. Visuals or graph tell us much more information than numbers can reveal. Most of the inferences can be established with this method. This step is the backbone of this project. Graph such as scatter plots, bar plots, box plots and heatmaps are used for this purpose.

Step 5: Hierarchical Clustering algorithms are applied on the variables to find the districts which fall under one cluster where a certain problem is prevalent. Moreover, what variables can be improved upon to bring these districts on par with those who are performing better.

The algorithm for hierarchical clustering is as follows:

1. Selecting the desired columns to form clusters.
2. Deciding the appropriate linkage for the clustering analysis and formation of dendrogram.
3. Selection of k (number of clusters to be formed) on the basis of dendrogram.
4. Fitting the dataset on the model.
5. Visualizing the clusters with the help of graph.

The project has been completed in Python language. The platform used for data cleaning, merging, Exploratory Data Analysis, Data Visualization and Clustering is Jupyter Notebooks.

Libraries used:

1. NumPy (Numerical processing tasks)
2. Pandas (Data pre-processing and transformation)
3. Matplotlib (Visualization library)
4. Seaborn (Visualization library)
5. Scikit learn (For Clustering Analysis)
6. SciPy – Scientific Python library (Used for Agglomerative Hierarchical Clustering)

6 Results

Final outcome of the project is always important in a data analysis project. The following pages will throw a light upon some of the findings achieved in this project.

In the Fig. 2, the proportion of working class increases as the sex ratio increases. This means that female also contribute to working population when they are good in number. The outlier here is Kerala, where sex ratio is good but the working population proportion is low.

The Fig. 3 shows the overall trend of child population versus working class population. In most of the cases, child population does not matter much. But the districts which are inside the red circle have low child population and at the same time have approximately half of the population under working class.

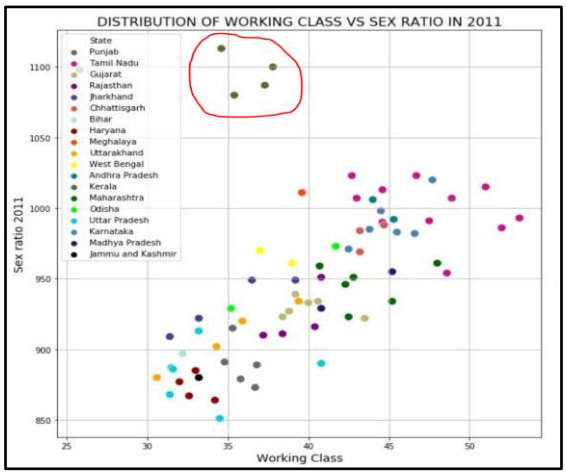


Fig 2. Distribution of working class and sex ratio in 2011

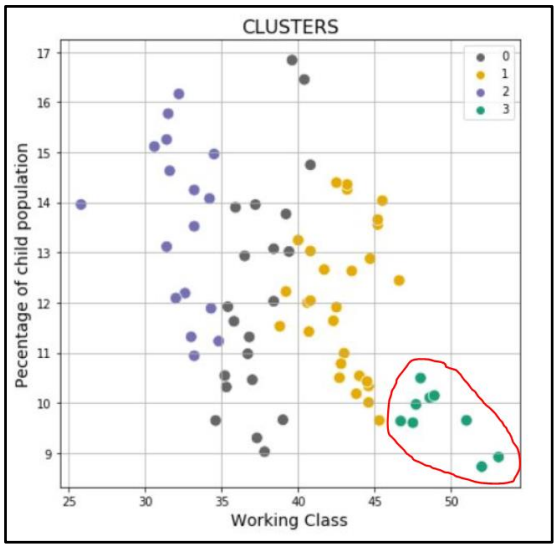


Fig 3. Clusters based on Child population percentage

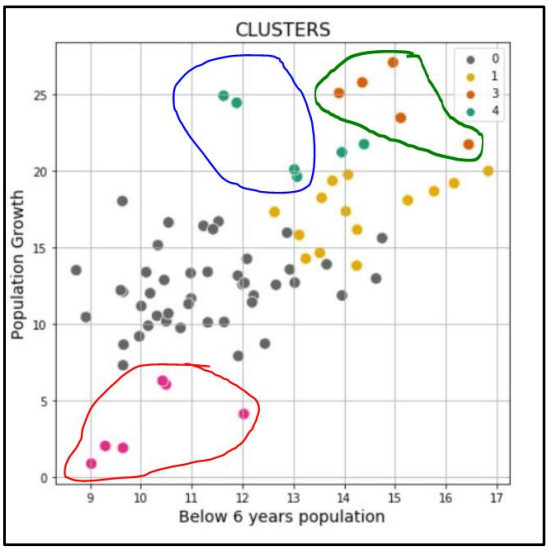


Fig 4. Child population with respect to population growth

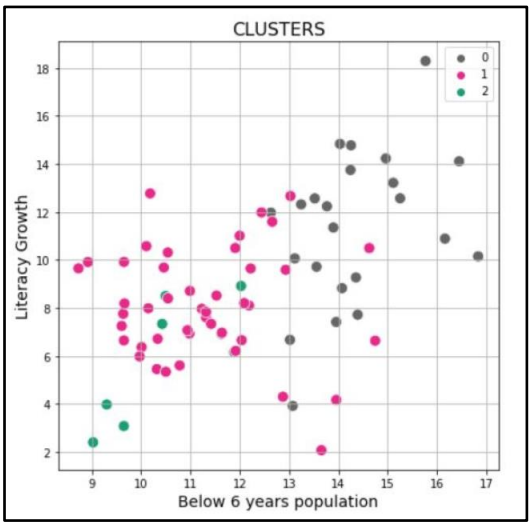


Fig 5. Child population with respect to literacy growth

In both the graphs, as the proportion of child population increases, so does the population growth and literacy growth. The districts inside the blue circle in the left graph have seen population growth without much increase in child population. To conclude, child population can increase overall population and can also lead to an increase in literacy.

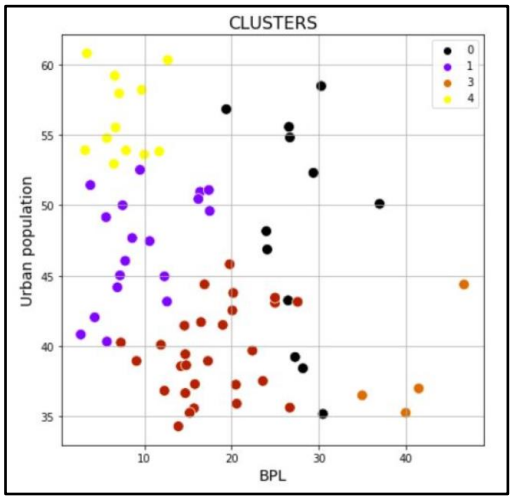


Fig 6. Relationship between urban population and Below poverty line population

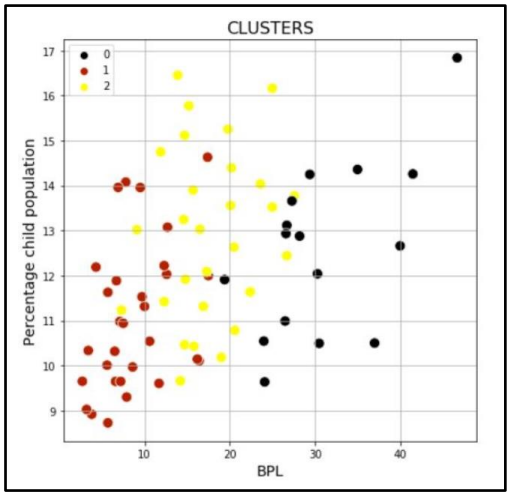


Fig 7. Relationship between Child population and Below Poverty Line

Overall, the trend shows that as proportion of people living in urban areas decreases, the percentage of people living below poverty line increases. Urbanization to some extent can help reduce poverty.

The graph on the right depicts the child population proportion to the below poverty line. It shows an increasing linear trend. More child population leads to more below poverty line population. A reason for this can be the difficulties faced by the previous generations to take proper care of children.

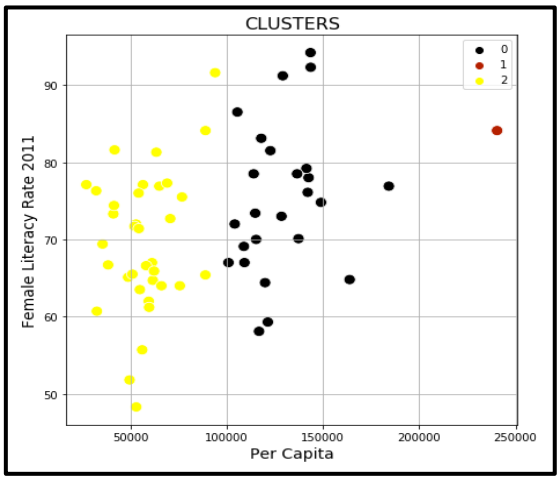


Fig 8. Female literacy rate 2011 vs per capita

From the scatter plot, it is clear that female literacy plays an important role to increase per capita income. When females work to earn a living along with men, the overall income of the family grows.

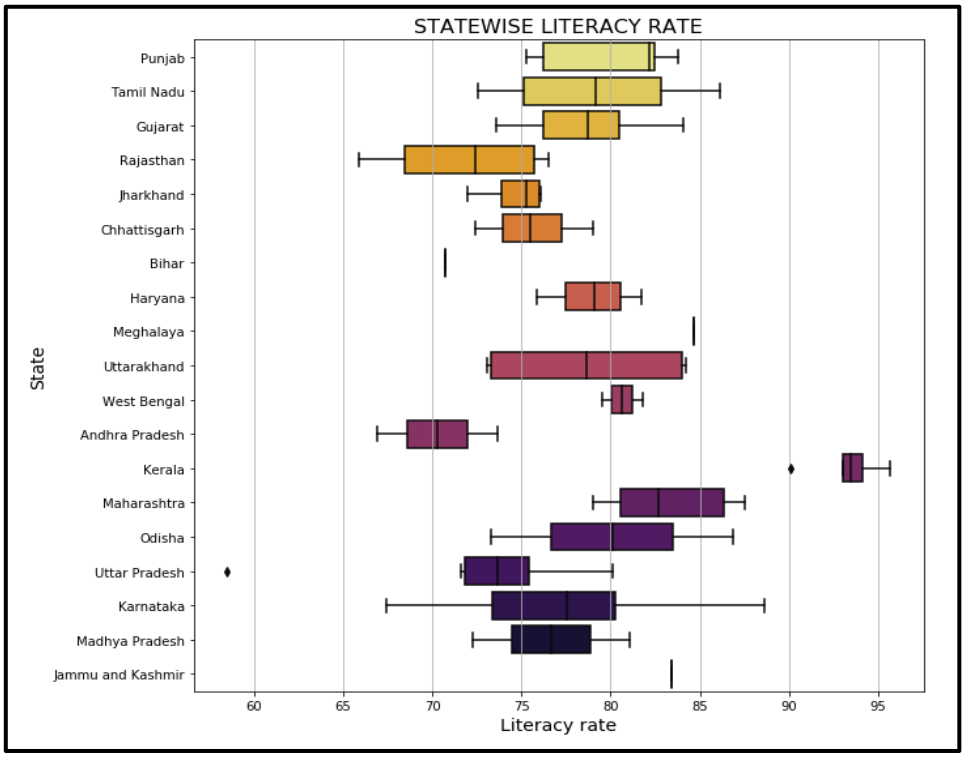


Fig 9. State wise literacy rates

Some districts of Rajasthan, Andhra Pradesh, Karnataka and Uttar Pradesh have literacy rates less than 70%. On the other hand, Kerala is the only state having literacy rate more than 90%. The states having low literacy levels are the ones having high infant mortality rates, low per capita income and more below poverty line population.

Some other important relationships are as follows:

1. As density of population increases, the proportion of population under working class decreases.
2. Infant Mortality Rate is lesser in areas where per capita income is high.
3. The same is true for Under five years mortality rate, places where per capita income is higher tend to have low Under five mortality rates.
4. High female literacy rates lead to greater per capita income.
5. High literacy rates lead to lower Infant Mortality and Under five years Mortality. This is true for male as well as female literacy rates.
6. Below Poverty Rate decreases as area under cultivation in a particular region increases.

The biggest factor that can bring about a change in the current scenario is the increase in literacy percentage. Male and female literacy can increase working class and thus help to increase average per capita income of the region. Increase in per capita income can help bring down infant mortality and under five mortality. It is rightly said that development can only be successful if the people of that area are literate. It has been observed in many districts that low literacy rate can act as an obstacle on the path to build an economically, socially, environmentally strong society.

7 Conclusion

A dream of making India a global superpower can only be achieved when not only the male child but also the female child gets equal opportunity to educate herself. Education alone can bring about a change in the current difficulties faced by the people of some semi-urban areas. The same solution may be valid for rural areas too which are further lagging in development than semi-urban areas. As future research, data can be gathered for air pollution, industries, crime rate, human development index, etc. These factors too play a pivotal role in the development of an area.

References

[1] https://www.collinsdictionary.com/dictionary/english/semiurban
[2] S. J. Meeus, H. Gulink, Semi-Urban areas in Landscape Research: A Review, Living Rev. Landscape Res., 2, 2008.
[3] https://www.census2011.co.in [Individual districts data]
[4] https://en.wikipedia.org/wiki/Telangana
[5] https://sangareddy.telangana.gov.in/about-district/
[6] https://en.wikipedia.org/wiki/Ranga_Reddy_district
[7] https://en.wikipedia.org/wiki/Bardhaman_district
[8] S. Ahuja, Indirect Estimates of District wise IMR and Under 5 Mortality using Census 2011 data, National Health Systems Resource Centre (NHSRC), New Delhi.
[9] https://www.livemint.com/topic/poverty%20grid [Individual states' article] (2014)
[10] https://agricoop.nic.in/, Department of Agriculture, Cooperation and Farmers Welfare. [Individual district reports]
[11] Central Ground Water Board Ministry of Water Resources, Government of India, Ground Water Quality in Shallow Aquifers of India, 2010.
[12] https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
[13] https://www.tableau.com/learn/articles/data-visualization
[14] https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
[15] https://uc-r.github.io/hc_clustering
[16] https://en.wikipedia.org/wiki/Complete-linkage_clustering
[17] https://en.wikipedia.org/wiki/Single-linkage_clustering
[18] R. Aijaz, India's Peri-Urban Regions: The Need for Policy and the Challenges of Governance, Observer Research Foundation, ORF Issue Brief No. 285, 2019.
[19] V. Narain, Peri urban water security in a context of urbanization and climate change: A review of concepts and relationships, Peri Urban Water Security Discussion Paper Series, Paper No. 1, SaciWATERs, 2010.
[20] F. J. Gallego, Mapping rural/urban areas from population density grids, Institute for Environment and Sustainability, JRC, Ispra, Italy.