Big Data Analysis of Education Attainment in Maharashtra State

Rupali D. Patil
Department of Statistics,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad- 431004 (M.S.) INDIA

Omprakash S. Jadhav
Department of Statistics,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad- 431004 (M.S.) INDIA

ABSTRACT
The census is the source of information about demography, literacy, education, fertility and mortality, religions etc. This treasure of information is beneficial for the government and planning commission. This overall combined census data is high dimensional and unstructured, so it can be classified as big data and the application of analysis of big data for the purpose of extracting patterns in several research fields is now a worldwide problem. Since education is pathway to any nation building enterprises, it created enlighten society, meritocratic human resources, democratic society etc. So, by considering key role of education in development socioeconomic growth, education attainment of Maharashtra state is considered for the study purpose. The districts of Maharashtra state with similar education levels are clustered using cluster analysis. The ST category showed higher variation in literacy as compared to other categories.

General Terms
Principle component analysis (PCA), Hierarchical cluster analysis, Census.

Keywords
Big Data, Data analysis, Chi-square test, Cluster analysis, Factor analysis.

1. INTRODUCTION
INDIA is a multi-cultural, multi-ethnic and multi-lingual country. There are three main religions, such as Hindus, Muslims and Christians in this country. The development of the country completely depends upon education. It is largely recognized as an important key towards a successful career for every individual worker to achieve a sustainable level of economic growth apart from formal education, the emphasis is more and more on lifelong learning and work related training activities. The role shaping and modifying the social structure is a topic of interest in building up of the nation. Moreover, women's education in India plays a vital role in improving life [9]. One of the particular concerns of education is the economic benefits deriving from various levels of education attainment. Most of the people believe in education as an ideal easy means of earning a high income. The people also believe that education is a key which opens up new opportunity for employment and provides more security and help to maintain the family and social status. A large number of empirical studies already dealt with the relationship between formal education and occupation. Apart from the economic growth in formal education also gives a proper guidance in finding the more knowledge about life skills and work-related training activities [4].

We tried to explore the relationship that exists between the educational attainment and the districts of Maharashtra state from the census survey 2011 of India. The educational levels according to category of the population like General Male literacy (GML), General female literacy (GFL), Schedule caste male literacy (SCML), Scheduled caste female literacy (SCFL), Schedules tribe male literacy and Schedules tribe female literacy (STML & STFL) are considered for study purpose.

1.1 Big Data and Census Data in India
Big data is the term for a collection of data sets which are large and complex; it contains structured and unstructured types of datasets. Data comes from everywhere; sensors used together climate information, posts to social media sites, digital picture and videos etc. which is classified as big data. The useful data can be separated from this big data with the help of advanced models and techniques of statistics. The statistical techniques play an important role in big data analysis. The statistical techniques used on unhidden pattern and uncovered information of big data extract into the useful information [2,7]. Indian census data are also the large sized data. The decennial census of India has been conducted 15 times, as of 2011. It has been conducted every 10 years, beginning in 1871 [13]. Hence, this huge data called as big data. For the analysis purpose a proper statistical techniques will be apply on this big data.

2. LITERATURE REVIEW
The concept of big data is just coming into existence and has uncertain origins. The term big data appeared for the first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of Big Data and NextWave of InfraStress. Currently, enormous amounts of data are created every day. With the rapid expansion of data, these data moving from the Terabyte to the Petabyte age [2,7].
Shariff Abusaleh reviewed the socio-economic and demographic data according to religion available from various censuses, National sample survey and academic publications since the independence of India. Discussed the structure and levels of employment, of living and of education, according to religion arc and also presented the fertility and mortality indicators, distribution and growth of population. This paper highlights the need to strengthen the data based which would allow a study of ethnic and religious differentials in socio-economic and education achievements [8].

Gouda Sateesh M and T.V.Sekher presented the differentials and factors associated with dropouts in India. Found that only 75 percent of the children in the age group 6 to 16 years were attending school. About 14 percent of the children never attended the school and 11 percent dropped out of school for various reasons. Observed that the dropout was high among the children belonging to Muslim, Scheduled Caste and Scheduled Tribes families. The dropouts among the children

belonging to illiterate parents were four times higher than that of the literate parents. It was also observed that if parents were not working, the possibility of dropout among their children was relatively high [3].

Panicker Remya previewed on what is big data, technologies used to build a big data infrastructure and how it can be useful in the development of the nation. In this paper, theoretical model is created that can be implemented or build in the future. It shows the opportunities that are useful in policy making and decision making with use of big data [5]. After that they also presented a systematic projecting of census data for nation using proper statistical tools of big data analysis. The census survey presents the demographic, social and economic status of country and available at planning office of government and is used for analysis purpose. It also discovers some suitable technologies that can be suited to develop such tool [6].

Mamatha M. et al examined the determinants of educational outcome in India. The study is based on the NFHS-III data of the representative sample from all over the country, to examine the relative impacts of social, economic and household on the likelihood of transition from one educational level to the next and found that because of the large population, Hindus stand first in either No-education or higher education where as Christians got better educational attainment than Muslims [4].

3. OBJECTIVES OF STUDY

The database for the present study is from the 15th national census survey. The data collected is so massive that all the issue are analysed properly for the valid result and hence, we make use of multivariate techniques in order to find out more fact from the data.

1) To check the literacy rate in rural and urban region of study area.
2) To examine category wise literacy of study area.
3) To find an association between education attainment and districts of Maharashtra state.
4) To find similarity of education attainment in all districts of Maharashtra state.
5) To find the category wise variation in literacy of study area.

4. RESEARCH METHODOLOGY

4.1 Data Source for the Study

For the present study, we considered the data from the 15th national census survey (2001-2010). The data gathered by the office of the register, general and census commissioner, India under the ministry of home affairs, government of India [10]. The Census is a special, wide-range activity, which takes place once a decade in India. Its purpose is to gather information about the general population, in order to present a full and reliable picture of the population in the country - its housing conditions and demographic, social and economic characteristics. The information gathered includes data on age, gender, country of origin, year of immigration, marital status, housing conditions, marriage, number of children, education, employment, travelling habits, etc. [12]. The 2011 Indian National Census has been conducted in 2 phases - house listing and population [13].

4.2 Methodology

The proper tools and techniques utilized in the census data can be achieved by extracting the judicious information.

Chi square test: The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

Cluster Analysis: Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

Factor Analysis: It is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. The factor analysis is a data reduction technique which removes the redundancy of data from a set a correlation variable. It represents the correlated variables to a smaller set of the derived variables and the factors that are formed are relatively independent of one another [11].

5. ANALYSIS

5.1 Distribution of Literacy

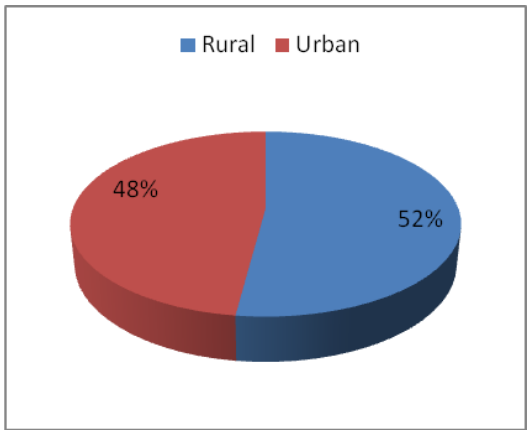

Fig 1: Distribution of literacy in rural and urban areas

We know that most of the population of Maharashtra state belongs to the rural background the present study represents the higher literacy rate in rural region.

5.2 Category Wise Distribution of Literacy rate in Maharashtra State

There are three major religions in Maharashtra state, namely Hindus, Muslims and Buddhists. We study, the category wise literacy rate in Maharashtra state.

Table 1. Category wise Distribution of literacy rate in Maharashtra state

Category	Frequency	Percentage	commutative %
STL	5887161	9.20	9.20
SCL	9285668	14.52	23.72
GL	48793114	76.28	100.00
Total	63965943	100	

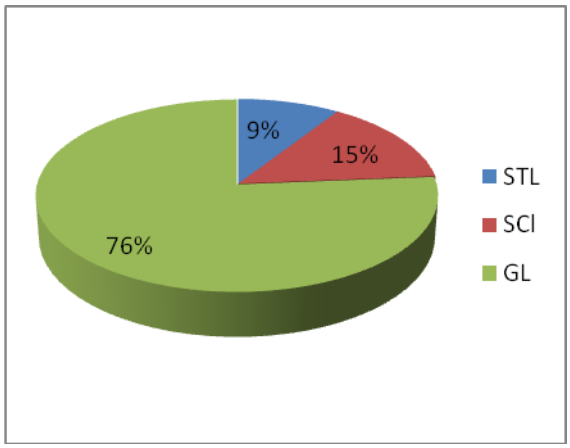

Fig 2: Category wise Distribution of literacy rate in Maharashtra state

From the above table , we observed that the literacy rate of ST population is very low in Maharashtra state.

5.3 Education Attainment

Education can be treated as an active process which is an individual's choice and each and every step of the process arises certain decision points or gates. In the present study, the education attainment consists of some education levels like literate without education level, below primary, primary, middle, metric, higher secondary, non technical diploma, technical diploma, graduate and above, unclassified. The following table gives the percentage of different levels of education in Maharashtra state**.**

Table 2. Percentages of Education Attainment in population of Maharashtra State

Graphics	Percentage	Cumulative %
Literate without education level (LWE)	1.97	1.97
Below primary (BP)	25.13	27.10
Primary (PRI)	25.4	52.55
Middle (MD)	14.91	67.45
Metric/secondary (SEC)	16.40	83.85
Higher secondary (HSEC)	7.76	91.61
Nontechnical diploma (NONTEC)	0.024	91.63
Technical diploma (TECH)	0.73	92.36
Graduate and above (GA)	7.6	100.00
Unclassified (UC)	0.001	100
Total	100	

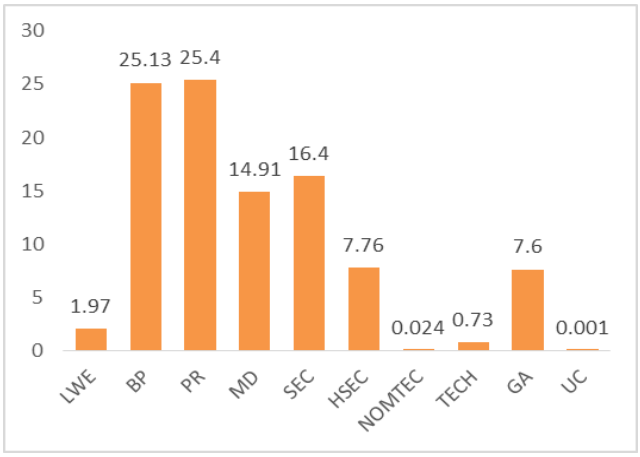

Fig 3: Education Attainment

5.4 Association between Education Attainments and Districts

The chi-square test to find the association between the two catogrical veriable from same population.

Ho: Education attainment of Population is independent of districts of Maharashtra state.

H1: Education attainment of Population is dependent of districts of Maharashtra state

Table 3. Pearson's Chisquare

	Value	D.F	P-value
Pearson,s Chi-square	2718178.562	306	.000

From chi-square test we conclude that there exists a significant relationship between population of education attainment and district of Maharashtra state.

5.5 Hierarchical Cluster Analysis of Education Attainment Acoording to Districts Maharashtra State

Hierarchical cluster analysis (HCA) is an exploratory tool designed to reveal natural grouping (or clusters) within a data set that would otherwise not be apparent.Hierarchical clustering analysis is computed for finding the similarities in education attainment population of different districts of Maharashtra states. At each stage in the procedure the number of groups is reduced by one, by joining together the group that is most similar and closer to each other. The visualization of such a hierarchical structure is a tree diagram generally known as Dendrogram.

HIRARCHICAL CLUSTER ANALYSIS (WARD METHOD)

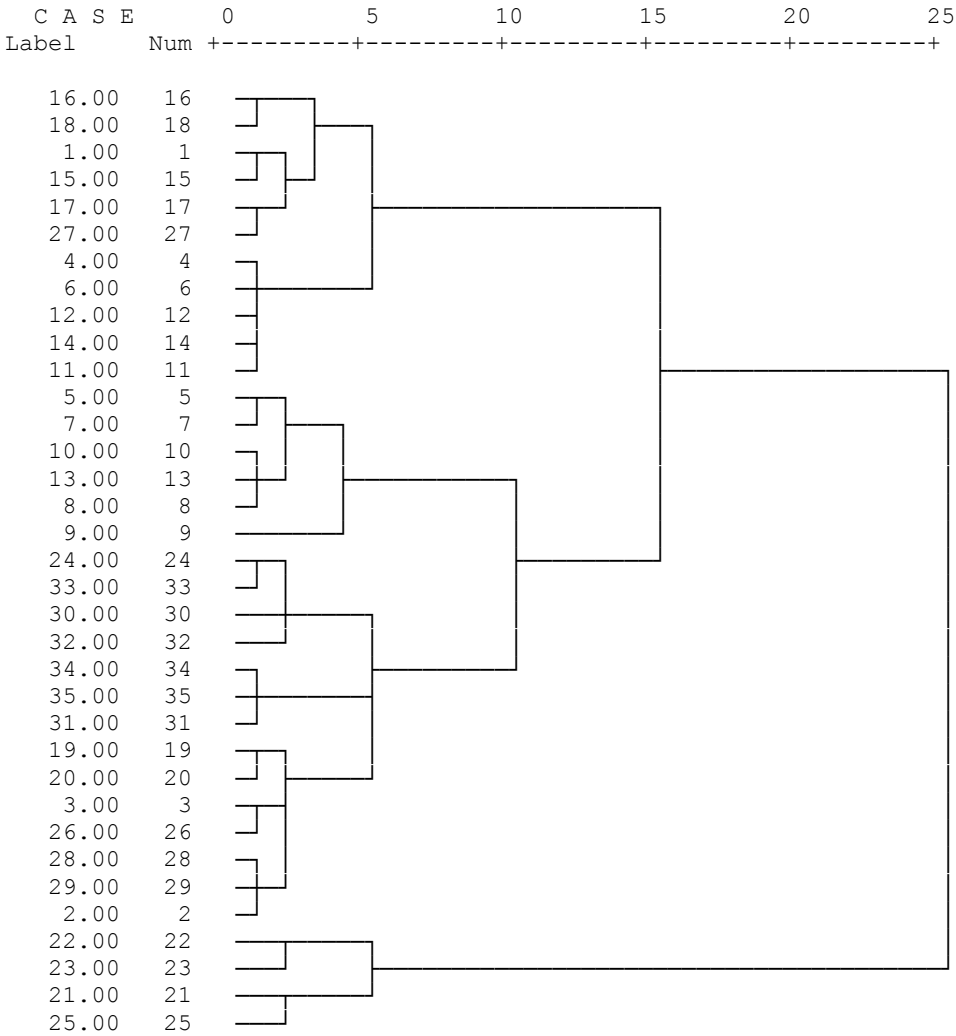

Fig 4: Dendrogram

The dendrogram of the clustering process shows the process of clustering pictorially about the relative similarity between cases. It is a tree diagram in which Y-axis represents the number of districts of Maharashtra state while the X-axis represents the distance. From the dendrogram, it is clear that six clusters with similar education levels of districts of Maharashtra are created . In first cluster 11 districts as Nandurbar, Buldhana, Washim, Gondiya, Gadchiroli, Yavatmal, Nanded, Hingoli, Parbhani, Jalna, Beed of Maharashtra stateare combined. The second cluster combines 9 districts such as Dhule, Jalgaon, Aurangabad, Nashik, Ahmadnagar, Latur, Osmanabad, Kolhapur, Sangli. The third cluster combins 5 district such as Akola, Amravati, Wardha, Nagpur, Bandra which has smaller distance between them. The fourth cluster consists of Thane and Pune districts. The fifth cluster have only two districts as Mumbai and Mumbai Sururban and the last sixth cluster create with Ratnagiri, Solapur, Raigad, Sindhudurg districts with minimum distance. These six clusters of districts of Maharashtra state observed high similarity between education attainment and the Nandurbar and Thane districts observe large dissimilarity between them.

5.6 Factor Analysis for Categorywise Literacy Rate in Maharashtra State

Factor analysis is used to reduce high dimensions data into much smaller size without losing the propertise of the data.

Table 4. Eigen Value for Literacy parameter

Principal Component	Eigen value	Total % of variance	Cumulative % of variance
First	3.95	65.89	65.89
Second	1.56	26.04	91.93

Table 5. Rotated Component Matrix

Literacy Parameter	Component	
	1	2
SC Male	**0.931**	0.125
SC Female	**0.928**	0.144
General Male	**0.917**	0.205
General Female	**0.913**	0.196
ST Male	0.175	**0.984**
ST Female	0.178	**0.984**

The ST male and female literacy exhibit a low degree of correlation among all the parameters and other parameters contribute high degree of correlation. Kaiser-Meyer-Olkin measure of sampling adequacy is 0.5 that indicate sample size is adequate for the application of factor analysis. The Bartlett's test of Sphericity has a significant value indicating the variable is non-normal. The Commonalities observed that ST male and female literacy have very high variations. From the total variance table, two components of variation have Eigen values greater than one and explain 65.89% and 26.03% respectively. From the rotated component matrix, 2 components extracted are

1. SC male, female and general male, female literate are strongly associated with factor 1.

2. ST male, female literates are strongly associated with factor 2.

From factor analysis, we observed that ST male and female population have low literacy as compare to other.

6. CONCLUSION

In Maharashtra state, literacy rate in rural region is 52% and in urban region it is 48%. In Maharashtra state only 7.6% population get higher educations. The cluster analysis of education attainment of district of Maharashtra state showed that Thane and Nandurbar district has a wide dissimilarity among their educational level.ST category showed higher variation in literacy rate as compared to SC and General literacy rate respectively. Overall analysis interpreted that; variation in education attainment is widely depending upon the development of district of Maharashtra state. The study of distribution of literacy according to the all district of Maharashtra state generally will be helpful to government and planning commission.

7. ACKNOWLEDGEMENTS

Authors are thankful to the office of the register, general and census commissioner, India under the ministry of home affairs, government of India for their valuable support.

8. REFERENCES

[1] Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl 2011. Cluster analysis, John Wiley and Sons, Ltd, 5th edition.

[2] Bharti Thakur, Manish Mann 2014. Data Mining for Big Data: A Review, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5.

[3] Gouda Sateesh M, Sekher T.V. 2014. Factors Leading to School Dropouts in India: An Analysis of National Family Health Survey-3 Data. Journal of Research & Method in Education, Volume 4, Issue 6, PP 75-83.

[4] Mamatha M. and Rao Nageswara V. 2015. Factors Influencing the Educational Attainment in India. International Journal of Mathematical Sciences, Technology and Humanities, Vol. 5, Issue 1, PP. 26-32.

[5] Panicker Remya 2013. Adoption of Big Data Technology for the Development of Developing Countries. Proceedings of National Conference on New Horizons in IT – NCNHIT, ISBN 978-93-82338-79-6.

[6] Panicker Remya 2016. Digitizing Indian Census Data for Analytics, Using Big Data Technology. International Journal of Advanced Research in Science, Engineering and Technology Vol. 3, Issue 3.

[7] Patil R. D., Jadhav Omprakash S. 2016. Some Contribution of Statistical Techniques in Big Data: A Review. International Journal on Recent and Innovation Trends in Computing and Communication, Vol.4, Issue 4, PP: 293 – 303.

[8] Shariff Abusaleh, 1995. Socioeconomic and Demographic Differentials between Hindus and Muslims in India. Economic and Political Weekly.

[9] Sujatha Sai D., Reddy Brahmananda G. 2009. Women's Education, Autonomy and Fertility Behavior. Asia – Pacific Journal of the Social Sciences, Vol. I, PP: 35-50.

[10] www.census.gov.in

[11] https://en.wikipedia.org/wiki/Factor_analysis

[12] www.cbs.gov.il/census

[13] https://en. wikipedia.org/wiki/2011_Census_of_India