Demographic Data Assessment using Novel 3DCCOM Spatial Hierarchical Clustering: A Case Study of Sonipat Block, Haryana

Mamta Malik[1]
Research Scholar, CSE. DCRUST
Sonepat, India
malik.mamta@gmail.com

Dr. Parvinder Singh[2]
Associate Prof., CSE, DCRUST
Sonepat, India
parvinder@rediffmail.com

Dr. A. K. Sharma[3]
Professor & Dean, CSE, YMCA University of Science & Technology
Faridabad, India
Ashokkale2@rediffmail.com

***Abstract*— Cluster detection is a tool employed by GIS scientists who specialize in the field of spatial analysis. This study employed a combination of GIS, RS and a novel 3DCCOM spatial data clustering algorithm to assess the rural demographic development strategies of Sonepat block, Haryana, India. This Study is undertaken in the rural and rural-based district in India to demonstrate the integration of village-level spatial and non-spatial data in GIS environment using Hierarchical Clustering. Spatial clusters of living standard parameters, including family members, male and female population, sex ratio, total male and female education ratio etc. The paper also envisages future development and usefulness of this community GIS, Spatial data clustering tool for grass-root level planning. Any data that shows geographic (spatial) variability can be subject to cluster analysis.**

Keywords- 3DCCOM(Three Dimensional Clustering with Constraints and Obstacle Modeling), Demographic Data; Geographical Information System (GIS); Hierarchical Clustering; Remote Sensing(RS)

I. INTRODUCTION

Rural development aims at improving the quality of life of rural people by utilizing natural and cultural resources available in rural areas, under the premises of existing legislature, in a sustainable manner through people participation. All the rural development [16], [17] programmes and schemes are still based on sectoral approach. But for a comprehensive rural development a comprehensive spatial development approach is required. For any successful rural development plan the interrelationship of farmland and population area must be understood carefully. So the study approach will be based on spatial totality (space, people and their activities) of rural areas in an integrated manner not in isolation of each other using hierarchical clustering technique[7], [11], [13]. The study will focus on the spatial database mining techniques to make utilisation of natural and cultural resources of rural areas in a sustainable manner with people's participation. The District Sonipat comes in to existence on December 22, 1972. It was a part of Rohtak district till December 21, 1972. The geographical location of Sonipat district is from 28° 48' to 29° 17' N, and 76° 28' to 77° 15' E. Sonipat district is one of the developed district of Haryana state as shown in figure1 and its economy mainly dependent upon agriculture. 74.87% of total population lives in rural areas and its main occupation is agriculture. In terms of agricultural production Sonipat district has surplus production but there is a lack of proper management of this surplus.

The standard paper components have been specified for two reasons: (1) Generating a spatial database [9], [14] of Sonepat Block for demographic assessment. (2) Use Hierarchical clustering technique [7] over spatial database to discovering knowledge [5], [9] that can be used for planning and sustainable development at grass root level.

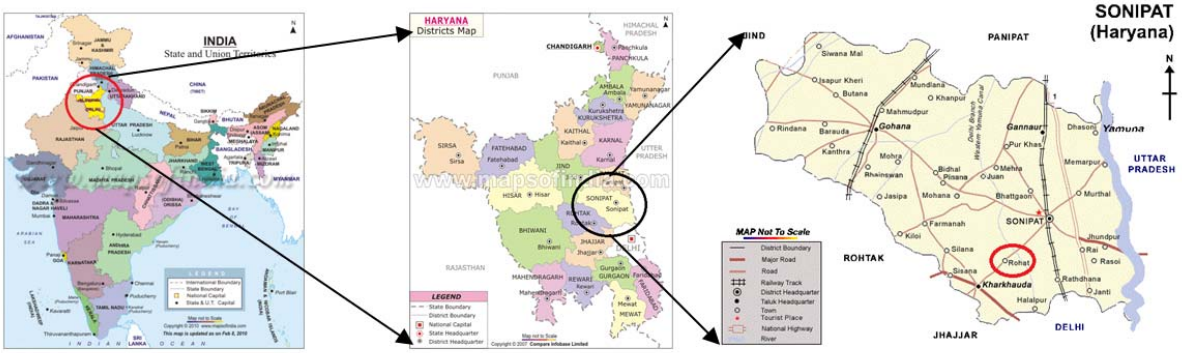


Figure 1. Graphical Location of Sonepat District, Haryana, India

Rural development [15] is the improvement in overall rural community conditions, including economic and other quality of life considerations such as the environment, health, infrastructure, and housing [16], [17]. For most small communities, this improvement involves population and employment growth, however, such growth is neither a necessary nor a sufficient condition for rural development. Approach for the clustering in rural areas has been clustered through 1) Identify and select indicators, 2) Use a correlation analysis, 3DCCOM hierarchical algorithm to choose number of clusters. In this study the two phenomena for which clustering were studied are block literacy and sex ratio of total population. The data was obtained from the census India website. The Block centroids were visually obtained by the authors using a GIS. The results were then displayed using a GIS tool ArcGIS and MapInfo [18].

II. METHODOLOGY AND DATA

The first law of Geography states that 'Everything is similar to everything else, but near things are more similar than farther things.' Statistically this property is defines as 'positive spatial autocorrelation'. Most Geographic phenomena tend to show positive spatial autocorrelation. Thus for instance, if the sonepat district of Haryana is found to have a high sex and literacy ratio, then we would expect its neighboring district to have same sex and literacy ratio rates more similar to the rate shown by sonepat than by let us say, a far away district like Sirsa. A 'cluster' would then be a collection of blocks, District with high literacy and sex rates. The question that now arises is, can a 'cluster' of blocks or district with high sex or literacy not be identified, by just drawing a chloropleth map displaying thesex or literacy rates? The answer is no that is because; a cluster can arise at random. Thus if the null hypothesis is considered as the case of no clustering, and if the null hypothesis is shot down and a cluster is found, then, the result has to be tested to check the probability of this result having arisen out of randomness. Thus the simple chloropleth map will not suffice. Thus we need a tool, into which, if the data (say literacy or sex rate) are fed, then it will come up with a cluster and carry out the necessary tests to check , if the resulting cluster has arisen out of randomness [19].

The tool used in this study is the spatial hierarchical clustering [1], [3]. We require multiple input files - one population file with male, female with their literacy rate, and one coordinate file for positional values. Thus, in our study the population file consists of the number of male, female literates in a particular block, say Sonepat, and thus the data is 'aggregated' at the block level in our case. The population file carries the population of the states, and the coordinate file contains the coordinates of the centroids of the blocks. 3DCCOM Spatial Hierarchical clustering working as by divisive whole state in multiple district and further district into block level. Once this testing is done, the test statistic is tested by generating a large number of random permutations of the dataset. Thus, the possibility of the cluster having arisen by chance is known. Depending on this individual block, we can then accept or throw away our results. It is also possible to output the results into a GIS to visually observe the cluster sizes. It must be noted that in the present context, the spatial scan statistic is being used as'Purely spatial' cluster detection. Clusters can also be detected in the temporal and the spatio-temporal domain. Such studies are ideally carried out with point level georeferenced data, but such data is totally unavailable in India and aggregated data has to suffice. Some commercially available data however, may have a better resolution than state level data.

A. Demographic Parameters

Analysis of demographic aspects is necessary for any type of study [17], [18]. These aspects tell about the structure and composition of society and findings of these aspects can be used for planning purposes. In this paper the following aspects of demography have been assessed at block level:

- Population Density;

- Sex Ratio;
- Literacy Rate; and

Rural development strategy [15] has been formulated on the basis of the issues identified through the analysis of primary and secondary data. Here rural development strategy has been formulated at sub-regional level. The problems identified at each sub-region level, are converted in to issues and then solution of these issues are defined in the form of strategy. Rural development strategy at sub-regional level for all the three sub-regions i.e. Most Developed, Medium Developed and Least Developed Sub-Regions, The whole district has been divided in to 3 sub-regions (Most Developed, Medium Developed and Least Developed) by analyzing the following parameters: population density, literacy rate and work participation ratio etc.

B. Hierarchical Clustering

Hierarchical clustering [1], [7], [11], [13], [14] creates a hierarchy of clusters which may be represented in a tree structure called a dendrogram as shown in figure 2 (a) and (b). The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations.

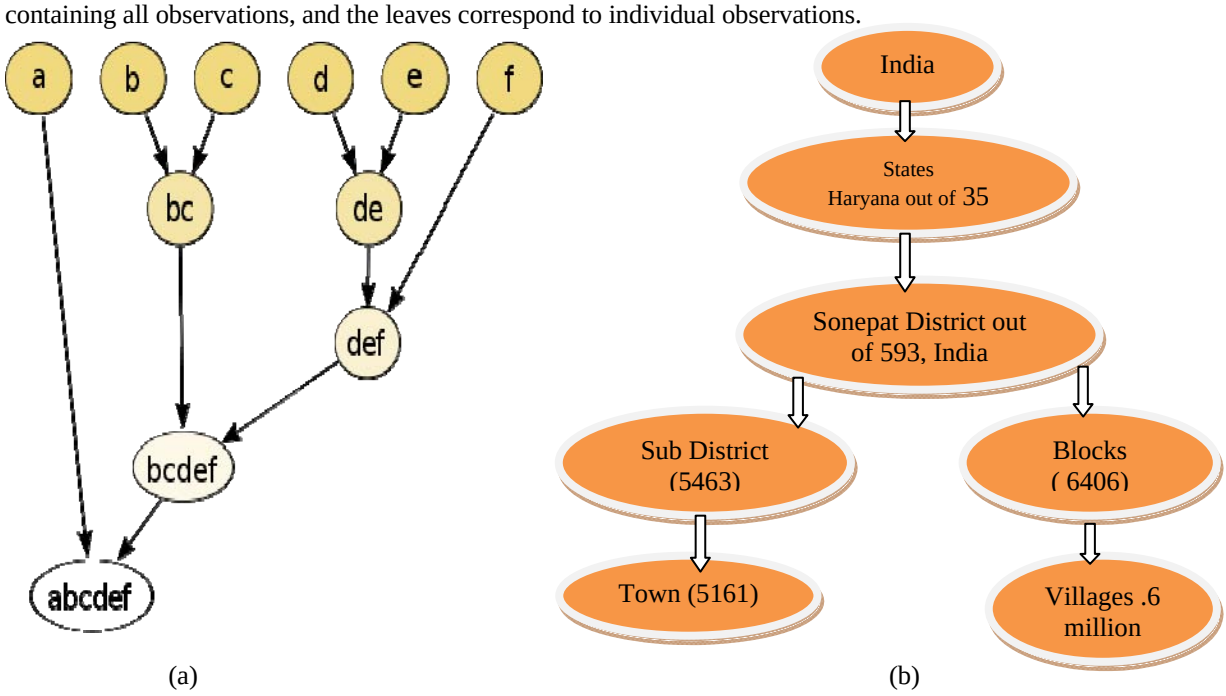


Figure 2. Dendrogram for hierarchical clustering (a) Agglomerative clustering (b) Divisive Country to village level clustering

Algorithms for hierarchical clustering are generally either agglomerative, in which one starts at the leaves and successively merges clusters together as in figure 2(a); or divisive in figure (b), in which one starts at the root and recursively splits the clusters. Any non-negative-valued function may be used as a measure of similarity between pairs of observations. The choice of which clusters to merge or split is determined by a linkage criterion, which is a function of the pair wise distances between observations. Cutting the tree at a given height will give a clustering [6], [8], [13] at a selected precision. In the following example, cutting after the second row will yield clusters {a} {b c} {d e} {f}. Cutting after the third row will yield clusters {a} {b c} {d e f}, which is a coarser clustering, with a smaller number of larger clusters. This method builds the hierarchy from the individual elements by progressively merging clusters. In our example, we have six elements {a} {b} {c} {d} {e} and {f}. The first step is to determine which elements to merge in a cluster. Usually, we want to take the two closest elements, according to the chosen distance.

III. A NOVAL 3DCCOM SPATIAL HIERARCHICAL ALGORITHM TO ACCESS DEMOGRAPHIC DATA

A. Hierarchical 3DCCOM Algorithm

The hierarchical version of 3DCCOM takes the concept of hierarchical clustering to produce a hierarchy of clusters [2], [3], [7]. It applies agglomerative approach of hierarchical clustering to a set of clusters produced by the 3DCCOM algorithm until a termination condition is satisfied. To better have an understanding of this, we have another definition to find the distance between the two clusters.

Let C = {c_1,c_2…c_k} be the set of clusters produced by any clustering algorithm. The distance between two clusters c_i and c_j is defined as:

dist (c_i, c_j) = Min{ dist(p,q) | p Є c_i and q Є c_j}.

The distance function takes all the points from two clusters and finds the distance, which is the distance between two nearest neighbors respectively from two clusters.

```
DBCCOM_Spatial_Heirarchical (Database D, obstacles O, Dmin, DT)
// Dmin is the minimum threshold distance between two clusters c1 and c2;
// DT is threshold density of every cluster
Dmin>Eps
Output: A hierarchy of Clusters.
C := DBCCOM(D,O);    //Start clustering
DO
FOR(D=DT ; D<= Dmax; D=D+DT)
FOR ( random ci and cj ε set of clusters C) do
IF (Dci ≤D) AND (Dcj≤D) AND (dist(ci,cj)≤Dmin ) // D is aligned threshold
density
Pts:=merge(ci,cj);
ClusterId:= assign_next_Id(pts,ClusterId);
Density of ClusterId=Dci+Dcj
Add ClusterId to C';
remove ci,cj from set of clusters C;
END IF;
ELSE
remove ci,cj from set of clusters C;
Add ci, cj to C';
END ELSE;
END FOR;
END FOR;
Write ClusterIds of clusters in C' to C;
WHILE (no more change from previous clustering);
RETURN hierarchy of clusters;
```

Algorithm 1: Hierarchical_3DCCOM

The threshold distance Dmin is taken to have an upper bound on the acceptable distance between two clusters. Where D_T is threshold density of clusters. Two clusters can be merged at a subsequent step if the distance between the two clusters is no more than Dmin and D_T should be threshold density of clusters to be merged into one cluster. The clusters at subsequent steps are merged together until the current clustering at a stage becomes similar to the clustering at the previous stage i.e. the number of clusters is same at two stages; this forms the termination condition for the hierarchical_3DCCOM() [13].

B. Demographic Data Assessment: A Case Study of Sonepat Block

Sonepat block from Sonepat district sub-region have been selected for detailed study. The parameters for selection of case study villages are same as the ones taken for sub-Regionalization of the district. These parameters are as follows:

- Population Density
- Sex Ratio
- Literacy Rate
- Work Participation Ratio

For selection of case study Sonepat Block, composite index is calculated for individual status and parameters are as shown in table 1 and 2. Ranks are then allotted to each parameter. Ranks of each parameter summed up horizontally. Finally categories have been allotted to each Sub-Block leading to selection Sonepat Block (one most developed and one least developed) from each sub-region as in table 3.

TABLE I. DEMOGRAPHIC STATUS OF ALL BLOCKS OF SONEPAT DISTRICT

Total Block of Sonepat District	Population Density per 1000	Sex Ratio per 1000	Literacy Rate (%)	Male Literacy Rate (%)	Female Literacy Rate (%)
Rai	596	811.69	75.58	85.97	62.83
Kharkhoda	542	834.14	69.36	80	56.74
Sonepat	1047	836.08	76.98	85.71	66.66
Ganaur	496	840.69	69.95	81.56	56.29
Gohana	566	878.82	71.92	82.46	60.15
Mundlana	346	840.8 5	67.27	79.21	53.17
Kathura	338	828.09	66.9	78.32	53.22

TABLE II. PARAMETRES FOR SONEPAT BLOCK

Name	Area in Sq. Kms.	Total Population	Male Population	Female Population	Density Per in Sq.Km	Sex Ratio	Literates	Literacy Rate	Scheduled Castes Population	Population % of SC
Sonepat	761.77	602896	329672	273224	791	829	392844	65	102157	17

TABLE III. MOST DEVELOPED AND LEAST DEVELOPED SUB-REGIONS OF SONEPAT BLOCK

Most Developed Sub-Region: Sonipat		**Medium Developed Sub-Region: Gohana**		**Least Developed Sub-Region: Mudlana**	
Most Developed	Least Developed	Most Developed	Least Developed	Most Developed	Least Developed
Blocks					
Rai	Ganaur	Gohana		Mudlana	Kathura

TABLE IV. DENSITY, LITERACY, SEX,WORK PARTICIPATIO RATIO OF SEVEN BLOCK OF SONEPAT DISTRICT BASED ON POPULATION

Item	Mundlana	Kathura	GOHANA	Ganaur	Sonipat	Rai	Kharkhoda
Density/Sq.Kms	1	1	2	2	3	2	2
Litracy Rate	3	3	1	2	1	1	2
Sex Ratio	2	3	1	2	2	3	2
Work Participation Ratio	2	1	3	2	3	1	1

The hierarchical version of 3DCCOM takes the concept of hierarchical clustering to produce a hierarchy of clusters. It applies agglomerative approach as shown in figure 3 of hierarchical clustering to a set of clusters produced by the 3DCCOM algorithm until a termination condition is satisfied. Agglomerative hierarchical clustering starts with initial stage by taking basic demographic parameters like, Total Male (M), Total Female (F), Male Literate (ML), and Female Literate (FL) for all the blocks of sonepat district. At the second stage of the clustering particularly based on distance and density parameter in proposed algorithm gives further assessment as total Male Ratio (MR), total Female Ratio (FR), Male Literate Rate (MLR), Female Literate Rate (FLR) etc. Similarly it will move until termination condition is satisfied as described in the 3DCCOM algorithm.

All the study Sub-Blocks have been assessed on the basis of primary and secondary data. The analysis of secondary data is necessary to know the spatial variation in terms of production, income, literacy rate etc as shown in table IV and primary survey is necessary to know the people's views and perception about the development. Village Information System (VIS) is a Geographical Information System (GIS) based application [18], [19], which provides detailed information pertaining to demography, infrastructure and natural resources for every village, district and state. But for a comprehensive rural development a comprehensive spatial development approach is required. For any successful rural development plan the interrelationship of farmland and population area must be understood carefully. So the study approach will be based on spatial totality (space, people and their activities) of rural areas in an integrated manner not in isolation of each other using hierarchical clustering technique. At the final stage algorithm gives best assessment and knowledge discovery within the basic demographic database like population status, literacy status, and work status for all level.

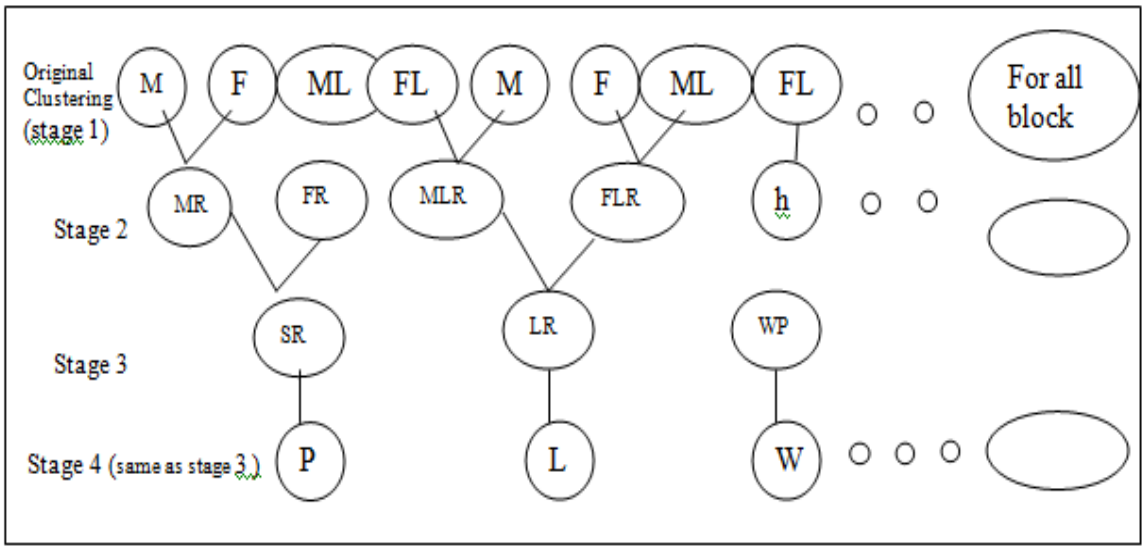
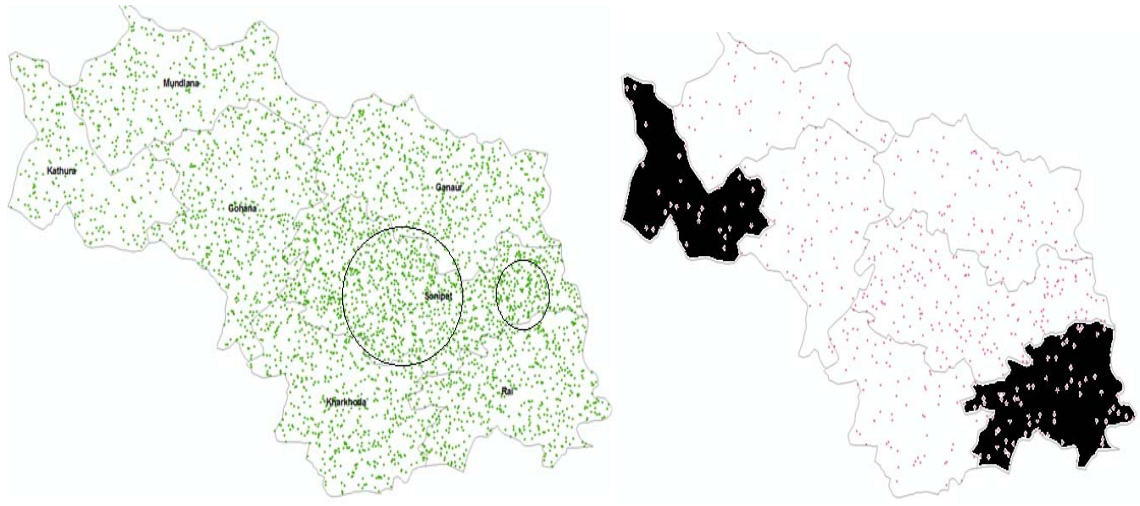


Figure 3. Agglomerative 3DCCOM Hierarchical Clustering for Demographic Assessment

The greatest challenge for sustainable development in the less developed world is to bring the scientific advances of the information technology and geographic information revolutions to bear on problems of severe environmental degradation while at the same time improving livelihoods. Although evaluating impact of information technology is complicated and difficult to carry out, we speculate on the impact or potential impact of geographic information technology on sustainable development. That's why technology plays a key role in planning process at every level. Agglomerative 3DCCOM hierarchical algorithm used to assess the demographic data. So that the results can be used for planning purpose at grass root level by keeping the idea of individual development.

IV. Experimental Results

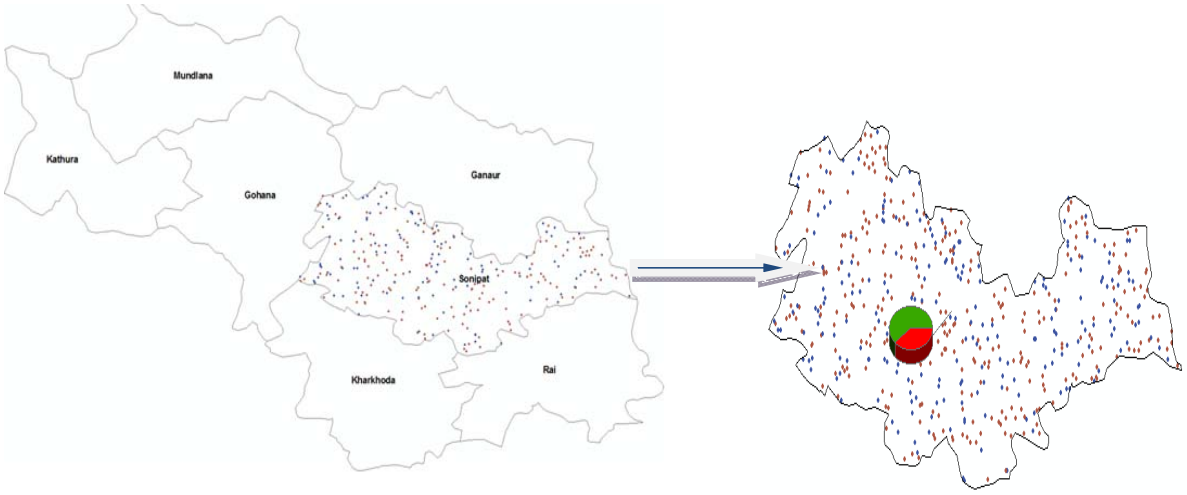
Often it is required to analyze the clusters at different levels. So, a hierarchy of clusters can be formed. By incorporating a hierarchical structure in the detected clusters, the overall search space can be pruned to a large extent and complexity at search levels can be reduced. Also the advantage obtained by reducing the search space dominates the complexity of the hierarchical algorithm. The overall results are significant in showing that clustering by using 3DCCOM algorithm. Spatial data can be represented easily in raster format i.e. data can be represented as points in n-dimensional space. Clustering analysis for data in a 3-dimensional space is considered spatial data mining and has applications in geographic information systems, pattern recognition, medical imaging, marketing analysis, weather forecasting, demographic data assessment etc as shown in figure 4(a), (b).



(a) (b)

Figure 4. (a) Highest population density cluster of sonepat block in sonepat district and (b) Least Population density with highest sex ratio of Rai and Kathura block of sonepat district at scale 1: 1000.

Figure 4 (a) Highest population density of sonepat block in sonepat district at the scale of 1:200 and figure 4 (b) Least Population density with highest sex ratio of Rai and Kathura block of sonepat district at scale 1: 1000 . Whereas figure 5(a) Male female literacy rate of sonepat block (b) Male female Literacy rate with pie diagram of sonepat block. Red color represents female literacy rate and Green color represents male literacy rate of total population at 1:1000 scale. Figure 6 describes a bar graph representes sex ratio of all sonepat district blocks with total population in yellow color, male density in red color and female population density in blue color.Maximum sex ratio is of kathura block in sonepat district.



(a) (b)

Figure 5. (a) Male female literacy rate of sonepat block (b) Male female Literacy rate with pie diagram of sonepat block. Red color represents female literacy rate and Green color represents male literacy rate of total population at 1:1000 scale

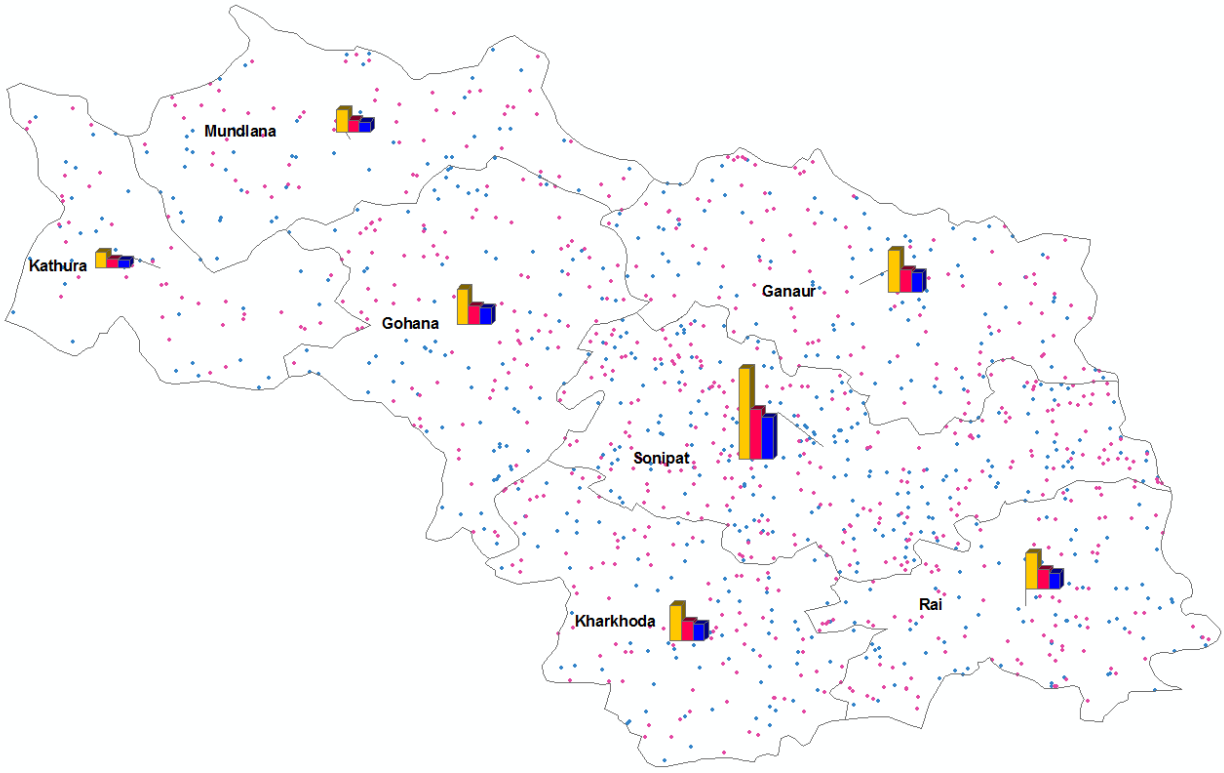


Figure 6. A bar graph representation of all sonepat district blocks with total population in yellow color, male density in red color and female population density in blue color.

Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard

analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign. For example, a company that sale a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving fewer sales then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process. To initiate the village development and planning process, a standardized and integrated Information System has to be developed. It should address the issues of micro level planning, integrated spatial development and horizontal as well as vertical integration of plans. Modern Geographic Sciences or Geo-spatial technologies (GIS, GPS and Remote Sensing) based Village Information System (VIS) can play a major role to achieve the desirable goal.

A. Advantages

Planning is now a widely accepted way to handle complex problems of resource allocation and decision-making. It involves the use of collective intelligence and foresight to chart direction, order, harmony and progress in public activity relating to human environment and general welfare. In order to provide a more effective and meaningful direction for better planning and development, necessary support of the organization has become essential. Hence the need for a suitable information system is increasingly being felt in all planning and developmental activities, whether these are for urban or rural areas. The position with regard to information system in urban areas however is far from satisfactory. Large volume of data is gathered whenever preparation of physical plan is taken up and a good number of maps as a part of the exercise on plan formulation are also prepared. However, today no system has been built to compile this geographically referenced data on a systematic manner and store them for retrieval at a subsequent point of time. Planners and decision-makers at Micro-level have to depend upon spatial and non-spatial data for optimal interpretation. Hence, the planners need to have at their disposal sophisticated data management systems to handle such spatially correlated data. The emergence of Remote Sensing and Geographic Information System as a powerful tool for spatial analysis and storage has in effect alleviated the problem by computerization of the spatial data. This new technology can reduce the time and cost to the planners in organizing the data in arriving at precise conclusion and decisions. It gives better results in assessment density based data e.g. demographic data assessment of sonepat block.

V. CONCLUSION

In this study the demographic phenomena for which clustering were studied are literacy, population, work and sex ratio. The data was obtained from the census India. The block centroids were visually obtained by the authors using GIS software. The results were then displayed using an ArcGIS and MapInfo. Results, conclusion and discussion Clusters of high literacy, population and sex ratio were obtained. Some sub-clusters were also obtained. The clusters for the two sets of data do not necessarily coincide. The results were then displayed in a GIS. Such results are of importance to national level policy makers and planners. To summarizes, cluster detection is a process of knowledge discovery. However, this is increasingly being used on various other data. In this study a cluster detection 3DCCOM algorithm is used on block level demographic data clusters identification.

ACKNOWLEDGMENT

The authors would like to thanks the DCRUST Sonepat and YMCA University, Faridabad, India to support us in research.

REFERENCES

[1] Han and M. KamberK. *Data Mining: Concepts and Techniques*. Morgan Kaufman, 2000.
[2] Ester M., Kriegel H.-P., Sander J., Xu X.: *"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise"*, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, 1996, pp. 226-231.
[3] Keim D. A.: "*Databases and Visualization*", Proc. Tutorial ACM SIGMOD Int. Conf. on management of Data, Montreal, Canada, 1996, p. 543.
[4] *"AUTOCLUST: Automatic clustering via boundary extraction for mining massive point-data sets":* Vladimir Estivill-Castro and Ickjai Lee Department of Computer Science & Software Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia.
[5] Osmar R. Zaïane: "Principles of Knowledge Discovery in Databases - Chapter 8: Data-Clustering" http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html.
[6] Beckmann N., Kriegel H.-P., Schneider R., Seeger B.: *"The R*-tree: An Efficient and Robust ccess Method for Points and Rectangles"*, Proc. ACM SIGMOD Int. Conf. On Management of Data, Atlantic City, NJ, ACM Press, New York, 1990, pp. 322-331.
[7] J G. Karypis, E. Han, and V. Kumar. Chameleon: *A hierarchical clustering algorithm using dynamic modeling*. In *IEEE Computer*, pages 68–75, 1999.

[8] P. S. Bradley, U. M. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In Knowledge Discovery and Data Mining, pages 9–15, 1998.
[9] Ester M., Kriegel H.-P., Xu X.: "*Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification*", Proc. 4th Int. Symp. on Large Spatial Databases, Portland, ME, 1995, in: Lecture Notes in Computer Science, Vol. 951, Springer, 1995, pp. 67-82.
[10] Hinneburg A., Keim D.: *"An Efficient Approach to Clustering in Large Multimedia Databases with Noise"*, Proc. 4th Int. Conf. on Knowledge Discovery & Data Mining, New York City, NY, 98.
[11] Jain A. K., Dubes R. C.: "*Algorithms for Clustering Data*," Prentice-Hall, 1988.
[12] Kaufman L., Rousseeuw P. J.: "*Finding Groups in Data: An Introduction to Cluster Analysis"*, John Wiley & Sons, 1990.
[13] Schikuta E.: "*Grid clustering: An efficient hierarchical clustering method for very large data sets*". 13th Int. Conf. on Pattern Recognition, Vol 2, 1996, p. 101-105.
[14] Ng R. T., Han J.: "*Efficient and Effective Clustering Methods for Spatial Data Mining"*, Proc. 20th Int. Conf. On Very Large Data Bases, Santiago, Chile, Morgan Kaufmann Publishers, San Francisco, CA, 1994, pp. 144-155.
[15] Whitener, Leslie. "Policy Options for a Changing Rural America," Amber Waves, Economic Research Service, USDA, April 2005. http://www.ers.usda.gov/Amberwaves/April05/Features/PolicyOptions.htm
[16] U.S. Department of Agriculture. Economic Research Service. "Nonmetro County Population Change, 2000-05." Amber Waves, June 2006. http://www.ers.usda.gov/AmberWaves/June06/Indicators/onthemap.htm
[17] U.S. Department of Agriculture. Economic Research Service. "Population, Income, Education and Employment, State Fact Sheets: United States." June 2006.
[18] Kumar, L & Kalyani, V.K. GIS AND ITS APPLICATION - RIT'2003
[19] Fischer, M., Scholten H.J. and Unwin, D. (eds.) 1996. Spatial Analytical.

AUTHORS PROFILE

Mamta Malik, First Author, is a research scholar, Department of Computer Science and Engineering, from DCRUST, Sonepat , India. Co-Authors are Dr Parvinder Singh working as Associate Professeor , Department of Computer Science and Engineering, from DCRUST, Sonepat , India and Dr. A. K. Sharma, Professor & Dean, Department of Computer Science and Engineering , YMCA University of Science & Technology, Faridabad, India.