

Method of Minimum Distance - A GIS Anchored System for Selection of Utility Service Stations

Anirban Chakraborty

Assistant Professor, Department of Computer Science

Barrackpore Rastraguru Surendranath College, Barrackpore, Kolkata – 120, W.B., India

Abstract - In this paper a GIS [4, 5, 6] anchored system has been proposed, enable to suggest the suitable locations for constructing the utility service stations. The information regarding the population of the customers in different wards/areas is fed through GUI in a digitized map [1], generally of a large region and to increase the portability of the system the information are kept using flat-file system. The technique takes as input the number of desired utility service stations to be constructed and then forms cluster of the adjacent wards/areas using the k-means method of clustering, under the condition that population of each cluster must be approximately equal. The proposed location of the utility service station and the areas covered up by any station are displayed graphically onto the map. The motivation behind selecting the location is that, the allotment of the customers to any service station should be done in a uniform way or in other words none of the customer should travel long to reach service station.

Key Words: Digitized map, flat-file-systems, ward, utility service station, clustering, K-means method of clustering

1.INTRODUCTION

Region segmentation plays a crucial role in GIS [4, 5, 6]. Segmentation of regions (generally done on input maps) is done to accommodate similar type of regions within one segment. If a map with many constituent regions are considered, then after segmentation, each segment will hold similar type of regions. Concept of map partitioning is almost same as region clustering. In most of the GIS anchored partitioning techniques, user targets to produce partitions in such a manner, so that either each partition will be of same area or each will hold same amount of attribute data, such as population.

The proposed segmentation technique works on the input digitized map, enriched with associated attribute data (whenever needed). This techniques work on the basis of the associated attribute data (here population has considered as the attribute data) of the fed map. The method chooses the sub-region possessing minimum centroid-to-centroid distance

with the centroid of the segment, for inclusion into that segment.

This proposed region segmentation [15] technique segments the fed input map on the basis of associated attribute data. For example, if population is considered as the attribute data, then this technique forms n segments, each of nearly equal population. The map taken into consideration for segmentation could be observed as composed of many smaller distinguished people domains (wards), for each of these wards the value of the considered associated attribute data (e.g. population) has fed. If each of these wards are considered as a separate node of a graph, then the entire map could be represented as a connected graph. Each region is identified by its unique identification number, generated automatically during digitization process. To demonstrate the technique of representing a map as a connected graph, let us consider a simple map constituted by five distinguished wards as represented in figure 1. The corresponding connected graph representation has been depicted in figure 2.

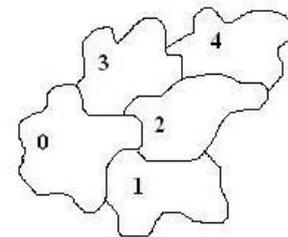


Fig. 1: A general map with five regions

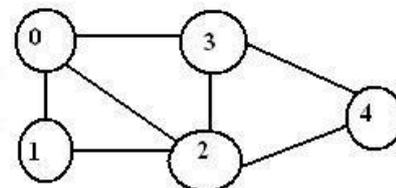


Fig. 2: Graphical representation of the map shown in figure 1.5

In a map two component regions are known to be adjacent of each other only when they share some common geographic boundary line. The adjacency information of the regions has to save using any dynamic data structure like linked-list. One such adjacency matrix representation has shown in table 1, with respect to figure 1 and 2. In this adjacency matrix, row and column indices signify the regions-ids and presence of a "1" in the cell(i,j) signifies that the regions i and j are adjacent.

Table 1: Structure showing adjacency of the regions (corresponding to Fig.)

	0	1	2	3	4
0	0	1	1	1	0
1	1	0	1	0	0
2	1	1	0	1	1
3	1	0	1	0	1
4	0	0	1	1	0

To achieve the objective of formation of segments of wards such that each segment should be populated with equal amount of associated attribute data; in this methodology formation of segments should have to meet the criteria that any segment should be formed by adjacent wards (regions) only and moreover the entire amount of associated attribute data considered should uniformly be distributed among these generated segments; so that finally a number of segments should be resulted; each of them of equal (approximately) size (in terms of associated attribute data, here population). It is quite obvious that at each step, the formation of segment is directed by the adjacency list, mentioned in table 1.

The organization of this paper is as follows. Section 2 of this paper deals with the proposed technique. The implemented results are given in Section 3. Analysis and Comparisons are outlined in Section 4 and finally Conclusions are drawn in Section 5.

2. METHODOLOGY

Using proposed Minimum Distance method of segmentation, the formation of segments could be achieved. The underlying idea is, given a data set of n regions; with x_1, x_2, \dots, x_n are the region centroids; such that each centroid point is in R^d , the problem of finding the minimum variance segmentation of the regions into k segments, is that of finding k points m_j ($j=1, 2, \dots, k$) in R^d such that

$$\frac{1}{n} \sum_{i=1}^n [\min_j d^2(x_i, m_j)] \dots\dots\dots (1)$$

is minimized, where $d(x_i, m_j)$ denotes the Euclidean distance between x_i and m_j . The points m_j ($j=1, 2, \dots, k$) are known as segment centroids. As here, the minimum distance between the region centroids and segment centroids are found, hence the name “Minimum Distance Method” has been adopted. The problem in equation 4.1 is to find k segment centroids, such that the average squared Euclidean distance (mean squared error, MSE) between a centroid of a region and its nearest segment centroid is minimized. Initially the border regions are taken (in a fashion— left-most, right-most, top-most, bottom-most and so on; this will help to chose initial regions to be apart from each other and will enhance the convergence speed of the procedure) as the initial cluster centroids and in each step the segments are expanded towards middle. If N is the entire population and k is the number of segments to form, then each segment will contain $\frac{N}{k}$ amount of population. As it

is not possible always to touch the exact figure of $\frac{N}{k}$, so a certain amount of threshold is considered. If it is T, then formation of a segment is announced to be complete, when its population attains the value $\frac{N}{k} \pm T$. The Minimum Distance Method segmentation procedure could be written in a step-wise manner, as depicted below.

Algorithm : Minimum Distance Method

- Step I. Digitize the map and populate it with associated attribute data
- Step II. Accept the number of segments to form, let it be k
- Step III. If N is the entire population and k is the required number of segments, then each segment will hold $\frac{N}{k} \pm T$ amount of population, where T is the deviation
- Step IV. Initialize the k border regions as initial segments in a fashion— left-most, right-most, top-most, bottom-most and so on
- Step V. At each step, each of k segments are expanded by inclusion of one region in it, provided the cluster is not complete and inclusion of the chosen region don't violate the threshold condition
- Step VI. Expansion of a segment is done by inclusion of a region (which is already not a member of any cluster) in it, for which segment's centroid and that particular region's centroid posses minimum distance.
- Step VII. Inclusion of region causes updation of the segment centroid at each step
- Step VIII. A segment is announced to be complete when it contains required amount of population.
- Step IX: End

Let there are 'm' constituent regions, which are to be segmented in 'k' parts. At each iteration, operations required are: calculation of distance between centroids, comparison between distances and calculation of modified centroid position. If the algorithm converges after 'I' iterations, then its time-complexity is $\approx O(Imk)$. For large data sets, where $k \ll m$, the time-complexity becomes $\approx O(m)$.

While implementing the above procedure of segmentation, a number of data structures have been used. It is the general requirement that any segment will be formed by only adjacent regions, i.e. regions having some common boundary line. Using the dynamic data structure linked-list, a table named “FinalTable” been implemented, which is used to hold the regions belonging to each segment. The row indices are the segment-ids and columns for any particular row contains the region-members of that segment. Initially, all the cells of this “Final Table” is made (-1). After selection of initial members of each segment (i.e. left-most, top-most, right-most etc. regions), the ids of the regions selected as initial member are placed into the first column of the “FinalTable”.

Another table named “Data Table” is there for holding data associated with each region. Here each row signifies a particular region (thus row indices are just the region-ids). The 1st column of any row signifies whether the region is already a member of any segment or not. Initially, for every row the value of this column is set to (-1), which signifies that the region is presently not a member of any segment. When any row (i.e. region) becomes the member of any segment, the

value is changed to (+1) and that region is not considered for any further steps of the procedure. The 2nd onward column(s) of each row of this table is used for storing data associated with the region.

As already stated, here the main objective of the work is to divide the entire population encountered among the segments, approximately equally. Thus if N is the total population and k is the number of segments formed, then each segment should hold approximately $\frac{N}{k}$ amount of population. A variable “sum” has introduced to hold this value $\frac{N}{k}$. If T is the deviation, then each segment will contain $sum \pm T$ amount of population. Using these data structures, the segmentation procedure mentioned in Algorithm 3 could be rewritten in a simplified manner as in Algorithm 4.

Algorithm Minimum Distance Method Procedure: Expressing another way

- Step I. for All the rows of “FinalTable” do
- Step II. For any row (if this segment is not complete), select all the column members
- Step III. All the adjacent regions of these column members are chosen and among these which is not member of any cluster and posses minimum distance with segment centroid is selected
- Step IV. if Due to inclusion of this region do not make the population of the cluster > (sum + deviation) then
- Step V. This region becomes a member of the segment
- Step VI. The segment centroid is updated accordingly
- Step VII. end if
- Step VIII. if The population of the segment is within the range (sum ± deviation) then
- Step IX. formation of this segment is complete
- Step X. end if
- 11: end for
- 12: Until there remains any un-segmented region, Go to step 1.

3. ILLUSTRATIONS

A step wise approach of implementation of the methodology in digitized map is explained illustrated using figure 3.

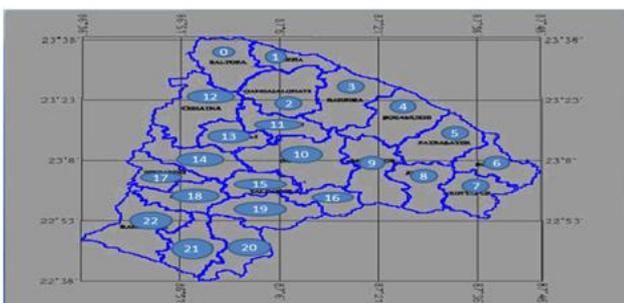


Fig. 3: Internally generated unique ID of the regions

Let us consider a digitized map of a very large area (Fig. 3), composed of 23 regions, identified by Region ids 0 to 22. These region ids are automatically generated during

digitization. Let the population of each individual region is as given in table 2.

Table 2: Population of the regions shown in figure 4.7

Region ID	Population
0	300
1	100
2	310
3	250
4	275
5	270
6	260
7	395
8	380
9	140
10	390
11	150
12	356
13	210
14	330
15	180
16	200
17	180
18	275
19	230
20	290
21	210
22	396

Thus the total population of the entire area considered is 6077. Let 3 segments has to form. The population of each segment will be approximately $6077 / 3 \approx 2025$, in order to make a uniform distribution of population among the segments. As it is not possible to touch the exact figure 2025 (because smaller constituent regions could not be sub divided), so an amount of deviation factor has been associated, depending upon the

application area. A deviation factor of amount 10% of the average population (rounded up to nearest multiple of 100), which is ≈ 300 for the present example, has been set. Thus the population of each segment should lie within the range $2025 \pm 300 = 1725$ to 2325 .

To form three segments for example, initially, left-most, top-most and right-most regions are selected, as the single members of each of the three segments. These are Region 22, 0 and 6. Thus now the population of Segment 1, 2 and 3 become 396, 300 and 260 respectively. In the 1st iteration, all the adjacencies of Region 22 are considered and the region possessing least centroid-to-centroid distance (and which is already not a member of any segment) is chosen. Thus Region 17 (with population 180) is selected as the new member of Segment 1. Applying same logic Segment 2 and 3 are also expanded. So, after 1st iteration the segments becomes as shown in table 4.

Table 4: Population of each segment after 1st iteration

Segment Name	Member Regions	Population
Segment 1	22,17	$396+180=576$
Segment 2	0,1	$300+100=400$
Segment 3	6,7	$260+395=655$

In the 2nd iteration all the adjacent regions of the present segment obtained after last iteration are considered and just like previous, region with minimum distance is selected. Thus, to expand Segment 1, all the adjacent regions of 22 and 17 (i.e. 14, 18 and 21) are considered and 18 is chosen (region with least distance). So, after 2nd iteration the segments become as shown in table 4.

Table 4: Population of each segment after 2nd iteration

Segment Name	Member Regions	Population
Segment 1	22,17,18	$396+180+275=851$
Segment 2	0,1,2	$300+100+310=710$
Segment 3	6,7,5	$260+395+270=925$

Proceeding in a same manner, finally the segments become as depicted in table 5.

Table 5: Population of each Segment after Final Iteration

Segment	Member Regions	Population
Segment 1	14,15,16,17,18,19,20,21,22	2291
Segment 2	0,1,2,3,10,11,12,13	2066
Segment 3	4,5,6,7,8,9	1720

Name		
Segment 1	14,15,16,17,18,19,20,21,22	2291
Segment 2	0,1,2,3,10,11,12,13	2066
Segment 3	4,5,6,7,8,9	1720

The centroid of each segment is also being pointed using red dots. A final pearance of this segmentation, with the centroid of each segment, is shown below in the figure 4.

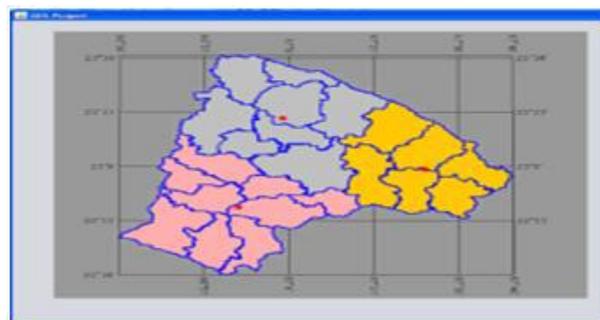


Fig. 4: Generation of three Segments by Minimum Distance Method

4. RESULTS

The implementation here is done by NetBeans (Java) [5] [6] and is based on flat-file systems, no usage of databases have been made to increase portability. Here it is foremost needed to insert a map of a very large area, say of entire sub-division or district. For the first time user, a New Profile has to be created with some meaningful name. For example, if the map of Kolkata, India is to consider (i.e. to deal with), with each of its wards as constituent regions, the profile may be adorned by the name “Kolkata”. Buttons for creation of new profile or working with existing profile are shown in figure 5.

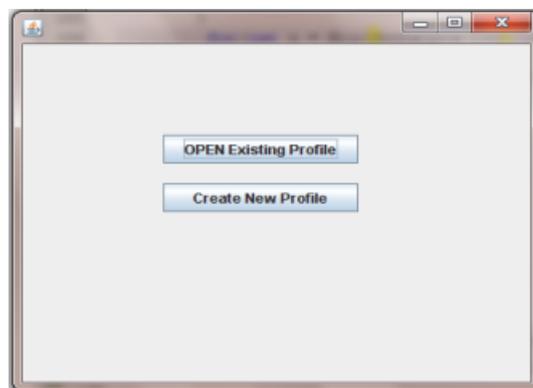


Fig. 5: Creation of New Profile

Digitization of raster map is done using any digitization tool. Figure 6 shows the GUI for data association.

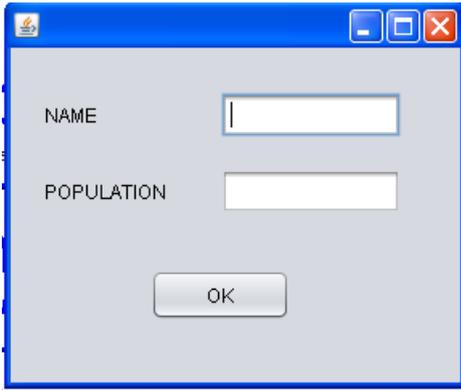
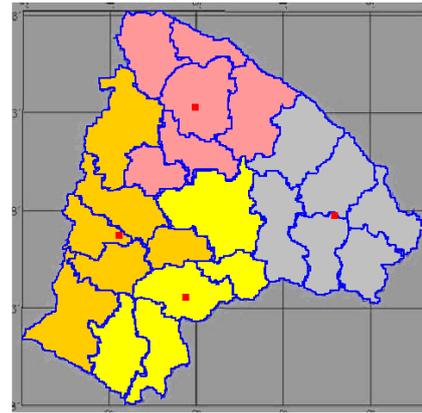


Fig. 6: Data Association in a vector map

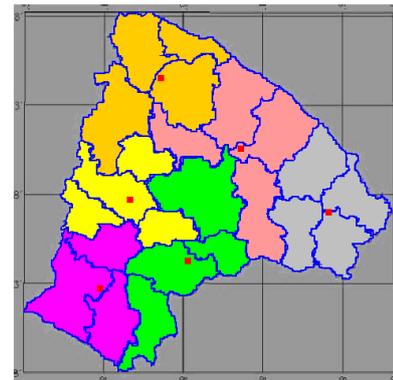


(c) 4 Segments generated

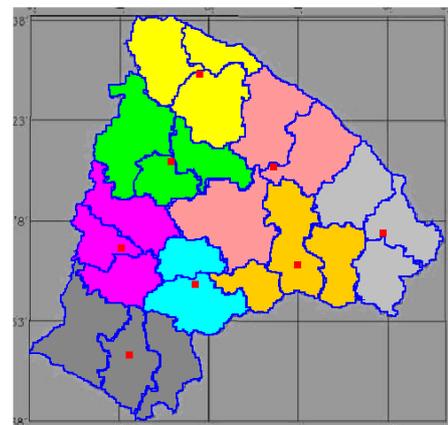
The formation of segments in Minimum Distance Method is illustrated in Figure 7.



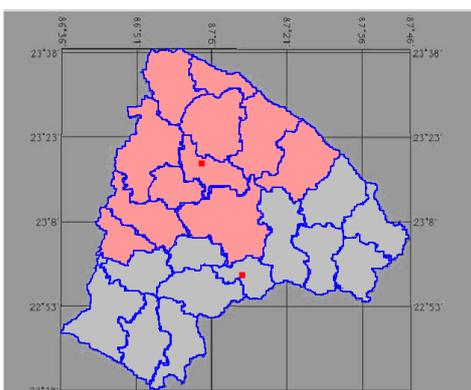
(a) Input map for segmentation



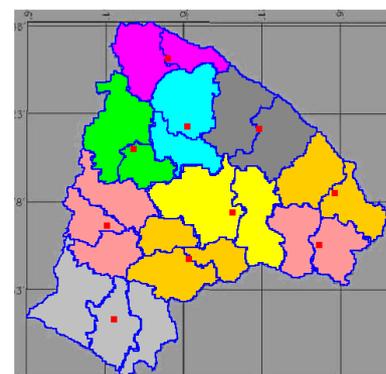
(d) 6 Segments generated



(e) 8 Segments generated



(b) 2 Segments generated



(f) 10 Segments generated

Fig. 7: Segments obtained by Minimum Distance Method

Figure 7 shows the result of five different number of segments, after applying the segmentation scheme on the input map shown in figure 7a. Table 7 shows execution time for the formation of five different segments (executed in a system, with specifications Dual Core Intel processor and 2 GB RAM).

Table 7: Execution time for formation five segments in Minimum Distance Method

Number of Segments	Execution Time (in milliseconds)
2	62
4	79
6	110
8	124
10	136

Figure 8 shows the graphical nature of time variation with number of segments.



Fig. 8: Graphical Representation of Number of Segments (Figure 7.5) vs. Time taken in milliseconds in Minimum Distance Method

5. CONCLUSIONS

For any country, especially for the rapidly developing third world countries like India, to meet the current need of its developing civilization, a numerous number of new constructions are needed. However the million dollar question is “Where”? To find the suitable most location for these new constructions is not easy enough. For any democratic country like India, election is supposed to be the backbone of the democracy. If Indian scenario is considered then it could be found that, taken into consideration different elections (like General or Lok Sabha, Assembly or Bidhansabha, Municipality, Panchayat etc.), election occurs almost once (sometimes even more) a year. Due to its high population, it is very much cumbersome and tedious to assign voters ward-wise into different polling station. The work becomes more

difficult due to the fact that, the voters should be assigned into different polling stations in a manner so that they have not to travel much. Otherwise, people will not be interested to cast their votes, which in turn will weaken the pillar of democracy. In spite of setting up of temporary utility stations like this, there are a lot of situations which demands same while constructing permanent utility service centers like setting up of new banks, health centers, electricity offices etc. The present technique is a solution of this problem, which automatically allocates customers into respective service stations in a uniform way, making the job fast and simple. Moreover, a graphical representation increases the readability of the output generated.

REFERENCES

- [1] 1. Anirban Chakraborty, Dr. J.K.Mondal, Arun Kumar Chakraborti, A File base GIS Anchored Information Retrieval Scheme (FBGISIRS) through Vectorization of Raster Map, International journal of advanced research in computer science, ISSN No. 0976-5697 volume-2 No.4, pp 132-138, July – August, 2011
- [2] Anirban Chakraborty, Dr. J. K. Mandal, N. Banerjee, P. Patra, A GIS based Interlinked Information Retrieval of a Large Database using KD Tree, Conference Proceedings “UGC-Sponsored National Symposium on Emerging Trends in Computer Science (ETCS 2012)”, ISBN number 978-81-921808-2-3, , pp 84-89, 20-21 January, 2012
- [3] J. K. Mandal, Anirban Chakraborty, Arun Kumar Chakraborti, A GIS Based Approach of Clustering for New Facility Areas in a Digitized Map, Conference Proceedings “International Conference on Eco-friendly Computing & Communication Systems (ICECCS 2012)”, CCIS 305, © Springer-Verlag, Berlin, Heidelberg, 2012, pp 398-405, 9-11 August, 2012
- [4] http://en.wikipedia.org/wiki/Geographic_information_system, accessed on 24 February, 2012
- [5] <http://www.gis.com/content/what-gis>, accessed on 24 February, 2012
- [6] <http://www.gisindia.in>, accessed on 22 March, 2012
- [7] <http://www.mapsofindia.com/gis-services.html>, accessed on 23 March, 2012
- [8] <http://www.gismaps.in>, accessed on 27 February, 2012
- [9] <http://www.maptell.com>, accessed on 25 February, 2012
- [10] <http://www.docs.oracle.com/javase/tutorial>, accessed on 12 May, 2012
- [11] <http://www.tutorialspoint.com/java/index.htm>, accessed on 14 May, 2012
- [12] <http://www.zetcode.com/tutorials/javaswing/tutorial>, accessed on 17 May, 2012
- [13] Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., “An efficient enhanced k-means clustering algorithm,” Journal of Zhejiang University Science A., pp. 1626–1633, 2006
- [14] Batty, M. and P. Longley, Analytical GIS: The Future, in Spatial Analysis: Modelling in a GIS Environment, P. Longley and M. Batty, Editors. 1996, Geoinformation International: Cambridge. p. 345-352.
- [15] Jain, A.K., M.N. Murty, and P.J. Flynn, Data Clustering: A review. ACM Computing Surveys, 1999. 31(3): p. 264-323.
- [16] Fayyad, U., et al., Advances in Knowledge Discovery and Data Mining. 1996: AAAI/MIT Press.
- [17] Han, J. and M. Kamber, *Data Mining : Concepts and Techniques*. 2000: Morgan Kaufmann.
- [18] Fukunaga, K., Introduction to statistical patterns recognition. 2nd ed. 1990: Academic Press Inc.
- [19] Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern Classification*. 2001: Wiley-Interscience.

- [20] Kaufman, L. and P. Rousseeuw, *Finding Groups in Data*. 1989: John Wiley and Sons.
- [21] Han, J., M. Kamber, and A. Tung, Spatial clustering methods in data mining, in *Geographic Data Mining and Knowledge Discovery*, H. Miller and J. Han, Editors. 2001, Taylor & Francis: London. p. 188-217.
- [22] Plane, D.A. and P.A. Rogerson, *The Geographical Analysis of Population: With Applications to Planning and Business*. 1994, New York: John Wiley & Sons.
- [23] Feng, Z. and R. Flowerdew, Fuzzy geodemographics: a contribution from fuzzy clustering methods, in *Innovations in GIS 5*, S. Carver, Editor. 1998, Taylor & Francis: London. p. 119-127.

BIOGRAPHIES



Dr. Anirban Chakraborty received his graduation degree with Physics Honours from University of Calcutta in 1999, completed his Masters Degree in Computer Application (MCA) from University of North Bengal in 2002, received his M.Phil. Degree in Computer Science from Madurai Kamaraj University in 2009 and received his Ph.D. from University of Kalyani in 2018. He has more than 20 years of teaching experience in College and University level. Presently he is working as Selection Grade Assistant Professor in Computer Science in Barrackpore Rastraguru Surendranath College, Kolkata-120.