Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Clustering of Indian districts based on supply chain requirements

A. Baskar

Panimalar Institute of Technology, Chennai 600 123, India

ARTICLE INFO

Article history:
Received 11 June 2020
Received in revised form 3 February 2021
Accepted 9 February 2021
Available online xxxx

Keywords:
Facility Location
Clustering
Fermat-Weber Problem
Weiszfeld's Algorithm

ABSTRACT

Facility location problems refer to the selection and placement of a facility to best meet the intended requirements. The problem often consists of fixing manufacturing premises, process industry or office location that minimises the total weighted distances between the data points and the selected centre. The weights may be the constraints or preferences among the data points. The solution should comply with the stated or implied constraints and maximise the profit. Distance is one of the important constraints that have a direct impact on supply chain costs. This paper considers the 661 districts of India (2011 census) as the data points and finds the centres by clustering the districts into the predefined number of clusters. These centres and number of districts attached to each centre vary depends on the supply chain requirements. We can assume that the main customers are located at the district headquarters and products are to be transmitted from the centres with minimum time and cost to these points. Different algorithms are used for fixing a facility; based on the population, based on the distance and so on. It is assumed that the districts' headquarters represent the entire districts. Geodetic coordinates are collected for these 661 districts and Haversine formulae are used for converting them into earth-centric earth fixed (ECEF) $\times$ x,*y* and *z* coordinates. Using these coordinates, the popular Weiszfeld's algorithm is used in addition to four other implementations to solve and find the clusters and total distance among the districts in each cluster. All algorithms are coded in MATLAB 2012a and run in an i5 PC with 4 GB RAM.

Selection and Peer-review under responsibility of the scientific committee of the International Mechanical Engineering Congress 2019: Materials Science.

1. Introduction

The well-known Fermat-Weber problem is one of the first problems in facility location theory [1]. This requires finding the 'geometric median of three points in a plane' assuming equal weightage (equal transportation cost) to all the points. Any facility location problem can be either a minisum or minimax problem. Fermat-Weber problem is a type of minisum problem wherein the sum of the weighted distances between the data points and the new facility is minimised. It can be mathematically expressed as:

Minimise

$$f(x) = \sum_{i=1}^{m} (w_i \| x_i - y \|_2) \quad (1)$$

$y \in \mathbb{R}^n$ where;
m – Number of data points
x_i – Data point
w_i – Weight for the data point x_i
y – New facility

The simplest model is when n = 2 and m = 3. That is, for the given three non-collinear points in a 2D plane, we have to find a fourth point such that the sum of its distances to the three given points is as small as possible. This problem was first geometrically solved by Torricelli in 1645. However, a direct iterative numerical solution was proposed by Kuhn and Kuenne in the year 1962 only, in the case of polygons having more than three sides [2]. The direct numerical solution was proposed by Tellier in 1972 for a triangular case [3]. The trigonometric solutions were analysed and generalised by Baskar for the problems with and without repulsion [4].

The original problem went through various modifications and implementations over time. For a weighted problem in '*n*' space and '*m*' data points, Weiszfeld's iterative algorithm finds an optimal solution by locating the 'geometric median' of the '*m*' points [5]. Chatzoglou et al. carried out exhaustive field research on the factors that have an impact on the selection of plant location [6]. Chen and Wang studied the facility location problems on the real line and their new algorithms break the ***O(nm)*** time bottleneck and solve the problems in sub-quadratic time [7].

E-mail address: a.baaskar@gmail.com

https://doi.org/10.1016/j.matpr.2021.02.292
2214-7853/
Selection and Peer-review under responsibility of the scientific committee of the International Mechanical Engineering Congress 2019: Materials Science.

Mahdian and Pál presented an approximation algorithm for the Universal Facility Location problem based on local search, under the assumption that the cost functions are non-decreasing [8]. A heuristic with application to ambulance location was proposed by Dzator and Dzator for the p-median problem to find the location of p-facilities so as to minimize the average weighted distance or time between demand points and service centres [9]. The heuristic uses a reduction and an exchange procedure and is effective for moderately sized problems.

However, not much literature is available about the clustering of places in a vast country like India for locating one or more similar facilities like establishing materials' supply points, a chain of manufacturing facilities of offices and, this work addresses the same.

2. Geography of India

India is a 'Unity in Diversity' country that has a lot of variations in culture, lifestyle, literacy level, food habits etc. The population as on today is estimated to be more than 136 crores, only next to China. In geography also, India has its differences. It has districts in states/ UTs that vary from 1 to 71 in number; population of 8004 to 11,060,148 persons in a district; population density from 1 person to 36,155 persons per square kilometre; area of an individual district from 9 to 45,674 square kilometres and the literacy rate varies from a minimum of 36.1% to a maximum of 97.91%. The state of Rajasthan has the largest state in India in terms of area with a share of 10.41% followed by Madhya Pradesh with 9.37%.

India is one of the largest countries by area in Asia, that measure:

North to South: 3214 km

East to West: 2933 km

Land frontier: 15200 km

Coastline: 7516.6 km

According to the 2011 census, there are 661 districts in India spread over 29 states and 7 union territories (UTs).

3. The K-Means clustering algorithm

Facility location or location analysis is concerned with the optimal placement of facilities to minimize the transportation or any other supply chain costs. Clustering of data points also forms a part of the solution in many cases. Clustering algorithms fall under unsupervised learning algorithms. The centres represent the clusters. Before starting the computation, the centres are assumed initially, either randomly or logically.

The data elements are assigned one by one to the nearest cluster CR, usually based on minimum distance. In each cluster, a new centre is computed to minimize the total sum of distances between CR and all other elements in the cluster. All elements are again checked and re-assigned to the nearest cluster.

The process is repeated until the algorithm satisfies the terminating condition.

4. Weiszfeld's algorithm

If a set of 'm' points $x_1, x_2, x_3, \ldots, x_m$; $x_i \in \mathbb{R}^n$ are considered, the geometric median is defined as:

arg min

$$\sum_{i=1}^{m} (\| x_i - y \|_2) \tag{2}$$

$y \in \mathbb{R}^n$

Here, "arg min" means the value of the argument 'y' that minimizes the sum. It is the point where the sum of all Euclidean distances to the 'x_i' is minimum.

Instead of only distance, the weighted distance also can be considered to obtain the weighted geometric median. In such cases, the sum of all the weights equals to 1.

Weiszfeld's algorithm is a form of iteratively re-weighted least squares. This algorithm estimates a new point from the existing point using the relationship:

$$y(i+1) = \frac{\sum_{j=1}^{m} \frac{x_j}{\|x_j - y_i\|}}{\sum_{j=1}^{m} \frac{1}{\|x_j - y_i\|}} \tag{3}$$

The algorithm converges for all initial points. Generally, the mass centre is taken as the initial point and the geometric median is reached after a finite number of iterations (Fig. 1).

5. Algorithms to identify the clusters

For identifying different clusters and analysis, five algorithms are considered, 'A' to 'E'.

A: Coordinates are randomly assigned for the initial centre. The new coordinates are taken as the average of the coordinates of the points assigned to the cluster, based on the distance. If the distance between the new and old centre is acceptable, the algorithm terminates.'y_i' refers to the present centre, '$y_{(i+1)}$' will be the new centre and each data point 'j' is represented by 'x_j'.

$$y(j+1) = \frac{\sum_{j=1}^{m} \| x_j - y_i \|}{m} \tag{4}$$

That is, the 'Mass Centre' estimated will be the cluster centre.

B: The new centre is based on the Weiszfeld's algorithm which moves the mass centre towards the 'Exact Centre' which is also known as the 'Geometric Median'.

$$y(i+1) = \frac{\sum_{j=1}^{m} \frac{x_j}{\|x_j - y_i\|}}{\sum_{j=1}^{m} \frac{1}{\|x_j - y_i\|}} \tag{5}$$

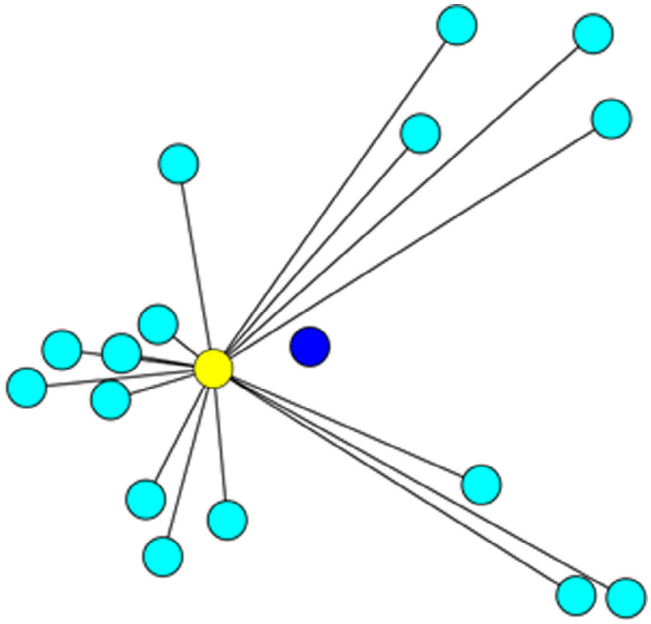

Fig. 1. The mass centre (in blue) moves towards the geometric median (in yellow) of a series of points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

C: This case considers only the '*x*' or '*y*' or '*z*' distance as the case may be, instead of Euclidean distance. For example, the '*x*' coordinate of the new centre will be estimated as:

$$x(i+1) = \frac{\sum_{j=1}^{m} \frac{x_j}{|x_j - x_i|}}{\sum_{j=1}^{m} \frac{1}{|x_j - x_i|}} \tag{6}$$

D: Similar to case B, Weiszfeld's algorithm is used here. However, the weight of each data point is the percentage population share, the sum of which equals to 1. That is, weighted Euclidean distance is used for the assignment of elements to any cluster.

$$y(i+1) = \frac{\sum_{j=1}^{m} \frac{w_j x_j}{||x_j - - y_i||}}{\sum_{j=1}^{m} \frac{w_j}{||x_j - - y_i||}} \tag{7}$$

E: This algorithm is similar to 'B'. However, it considers the Manhattan distance instead of Euclidean distance. It is the L_1 norm and can be represented as:

$$|x_j - x_i| + |y_j - y_i| + |(z_j - z_i)|$$

$$y(i+1) = \frac{\sum_{j=1}^{m} \frac{x_j}{|x_j - x_i| + |y_j - y_i| + |z_j - z_i|}}{\sum_{j=1}^{m} \frac{1}{|x_j - x_i| + |y_j - y_i| + |z_j - z_i|}} \tag{8}$$

Except for the case 'D', the weight of each data point is taken as 1. Also, only the Euclidean distances (L_2 norms) are considered except for the cases 'C' and 'E'.

6. Methodology

The latitudes, longitudes and altitudes of the all 661 districts headquarters are collected mostly from a single source to have consistency [10]. These geodetic coordinates are converted into earth-centric earth fixed (ECEF) × *x*, *y* and *z* coordinates. The ECEF is a geocentric Cartesian coordinate system having its origin (0,0,0) at the Earth's mass centre. The clusters are formed using these ECEF coordinates.

Haversine formulae are used to calculate the distance between two points of known latitudes, longitudes and altitudes [11]. The cluster centres obtained in ECEF coordinates are re-converted to geodetic coordinates from which the locations are identified [12,13]. Haversine formulae are simple to understand and code and are accurate to around 0.3%, which is still good enough for most applications. Hence, Haversine formulae are used in this work.

The earth's latitudes are parallel to each other and hence, the distance computed between each degree almost remains constant throughout. On the other hand, as the earth is slightly elliptical, minor variation between the degrees of longitudes is evident if we move away from the equator towards the poles.

- Two consecutive degrees of latitude are approximately111 kilometres apart.
- At the equator, the distance is 110.567 km.
- At the poles, the distance is 111.699 km.

In the case of longitude, the distance between degrees varies greatly. They are farthest apart at the equator and converge at the poles.

- A degree of longitude is widest at the equator with a distance of 111.321 km.
- The distance gradually shrinks to zero as they meet at the poles.

6.1. Conversion from geodetic coordinates to ECEF coordinates

The following formulae are used in the conversion process:

x, *y* and *z* - ECEF coordinates

a – Equatorial earth radius = 6366710 m as per World Geodetic System, 1984 (WGS84)

$C = \frac{1}{\sqrt{Cos^2(Latitude) + (1-f)^2 x Sin^2(Latitude)}}$; $S = (1\text{-}f)^2 \times C$

h – Altitude above the reference

f – Flattening parameter = (*a*-*b*)/*a* ... [*b* – Polar earth radius and, (*1*/*f*) = 298.257224]

x = (a.C + h).Cos(latitude).Cos(longitude)

y = (a.C + h).Cos(latitude).Sin(longitude)

z = (a.S + h).Sin(latitude).

6.2. Conversion from ECEF coordinates to Latitude, longitude and altitude

For converting back the ECEF coordinates to geodetic coordinates, another set of formulae are used:

longitude = $\tan^{-1}2(y/x)$

latitude = $\tan^{-1}2((z + ep^2.b.Sin(th)^3)/(p\text{-}e^2.a.Cos(th)^3))$

altitude = p/Cos(latitude)-N where,

$N = a/\sqrt{(1 - e^2).Sin(latitude)^2}$

Eccentricity, $e = \sqrt{2f - f^2} = 0.0081819190842622$

The constants are: $ep = \sqrt{(a^2 - b^2)/b^2}$; $p = \sqrt{(x^2 + y^2)}$; $th = \tan^{-1}2(a.z/b.p)$.

6.3. Distance between two points

For calculating the distance between two points of known geodetic coordinates, Haversine formulae are used. Using the formulae, they are converted into the ECEF coordinates. Now, using the conventional formula the distance can be computed.

Euclidean distance between two points with ECEF coordinates (x_1,y_1,z_1) and (x_2,y_2,z_2)

$= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$.

7. Computational results and discussion

All the algorithms are coded in MATLAB 2012a and run in an i5 PC with 4 GB RAM. The codes are written based on the references available in the Department of Commerce, US Govt. website [14]. The results obtained are validated using the calculator available in that web site and are perfectly matching.

Besides, two more problems considering the weights for each point used by Cooper and Katz were analysed [15]. The results obtained perfectly match with Cooper and Katz both in terms of magnitudes and number of iterations.

The number of clusters considered is from 1 to 5. Four out of five algorithms consider just the distances whereas, the algorithm 'D' uses (population) weighted distances for clustering the data points. If k = 1 (one cluster), we can assume the case that an organization is looking for establishing a single facility in India. It can be located in the cluster centre. For other values of '*k*' (2 to 5), the cases may be establishing '*k*' facilities across the country minimizing the total distance connecting the data points (districts headquarters) attached to a specific cluster centre, as well as the total distance. If any organization is interested more on the population rather than the distance (eg. a toothpaste company), the (population) weighted distances are considered instead of only distances. In such case, the locations of the cluster centres move towards the most populated points increasing the total distance.

For the analysis purpose, in addition to Euclidean distance; '*x*', '*y*' and '*z*' distances and Manhattan distances are also considered to find whether they can result in better solutions or not when compared to Weiszfeld's algorithm. Great Circle Distance (GCD)

Table 2
Number of Clusters = 5.

Cluster Type	No. of Points in each Cluster	Latitude	Longitude	Distance in each Cluster	Total Distance
A	94	25.7021747	92.6313333	24283.95	217920.59
	107	22.6378420	74.42345	37903.58	
	147	30.2093840	76.9132184	44391.15	
	134	14.1280929	77.8503002	51750.75	
	179	23.7292881	84.0419838	59591.16	
B	92	25.9225357	92.8654027	23205.00	215971.09
	114	22.2808444	75.0077237	42044.89	
	154	29.5028070	77.0351362	46913.62	
	126	13.4421182	77.6739726	46879.63	
	175	24.0729236	84.4250975	56927.95	
C	92	26.1688885	92.9748099	23332.90	216868.41
	112	22.6452429	74.91355597	40496.55	
	147	29.6565295	76.9343152	43830.06	
	134	14.25108589	77.90855385	52046.60	
	176	24.12419715	84.08486369	57162.30	
D	92	26.1365888	91.7576150	24765.01	221987.58
	120	20.8809557	74.6831757	47761.98	
	174	28.5283711	77.4303230	58339.95	
	111	12.9453872	77.8629813	38772.83	
	164	24.6530118	85.1065743	52347.81	
E	92	25.9004232	92.7767420	23215.31	215989.85
	114	22.3205153	74.9774895	46914.01	
	154	29.5102829	77.0457544	46880.67	
	126	13.4194426	77.6698450	46880.67	
	175	24.0487884	84.4598936	56931.38	

Total distance varies by a maximum of: (221987.58–215971.09) = 6016.49 km or 2.786%.

Table 3
Location of Cluster Centres, Total Distance and Variation in Total Distance between Extreme Values, % (Algorithm 'B').

No. of Clusters	Location of Cluster Centre(s)	Total Distance, km	Variation in Total Distance Between Extreme Values, %
1	Deora Khurd, Madhya Pradesh	548611.37	1.872
2	Keroli, Karnataka Shahpur, Uttar Pradesh	415607.28	0.671
3	Bich Maqo, Jharkhand Bandhwari, Haryana Pusalpahad, Telangana	300012.05	1.926
4	Banjhikend, Jharkhand Jajjal, Haryana Kesthu, Tamil Nadu Sonori, Maharashtra	259746.62	3.263
5	Kurkut, Assam Lakhankot, Madhya Pradesh Karnal, Haryana Jadalathimmanahalli, KarnatakaPakariya, Jharkhand	215971.09	2.786

is more accurate than ED when the data points lie on the earth surface. However for shorter distances, the difference is not significant. Also, when the number of clusters increases; the distances between the extreme points and centre decrease. Hence, GCD is not considered in this work.

In the real supply chain, the road distance will not be equal to ED or Manhatten or GCD in actual cases. Wiggle factor is a correction factor defined as the ratio between the real distance travelled by road and the straight line (or) aerial distance between the two points. It is used to estimate the route distances for road transport and subsequently the actual fuel cost. Generally the distance obtained is multiplied by a Wiggle factor of 1.2 to 1.25 which may vary from country to country.

The outputs obtained are analysed for clusters 1 to 5. The results for one cluster and five clusters are listed in Tables 1 and 2 for reference. The variations in the total distance are also given. The variations are observed to be from 0.6707% (for two clusters) to 3.2629% (for four clusters). The number of districts attached to a cluster and their respective centres is not consistent. The number of iterations before the results converging to the required accuracy level also differs. The algorithms terminate if the difference in the total distance between any two consecutive iterations is close to zero. If the algorithms are ranked by ordering them for the total distance we get:

One Cluster: B-E-A-D-C; Two Clusters: B-C-E-D-A

Three Clusters: B-E-C-A-D; Four Clusters: B-E-C-A-D

Table 1
Number of Clusters = 1.

Cluster Type	No. of Points in the Cluster	Latitude	Longitude	Total Distance in each Cluster
A	661	23.4405026	80.8410316	549531.52
B	661	23.9406761	80.7788587	548611.37
C	661	24.4129616	79.0083600	558881.31
D	661	23.5240091	80.310167	549912.16
E	661	23.9269927	81.182942	549084.46

Total distance varies by a maximum of (558881.31–548611.37) = 10269.940 km (1.872%).
Similarly,
For 2 Clusters: Total distance varies by a maximum of: (418394.56–415607.28) = 2787.28 km or 0.6707%.
For 3 Clusters: Total distance varies by a maximum of: (305789.42–300012.05) = 5777.37 km or 1.9257%.
For 4 Clusters: Total distance varies by a maximum of (268222.01–259746.62) = 8475.39 km or 3.2629%

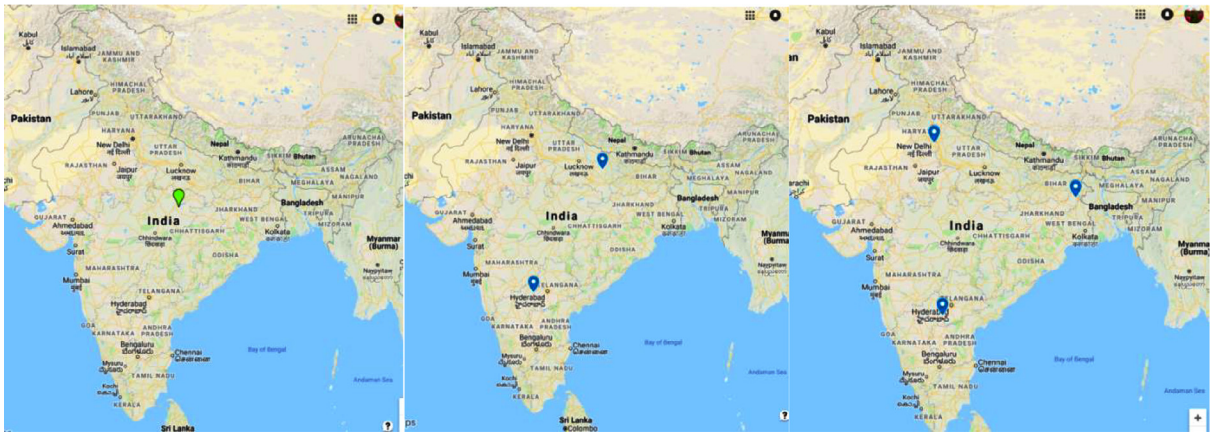

Fig. 2. One to Three Clusters and their Centres.

Fig. 3. Four to Five Clusters and their Centres.

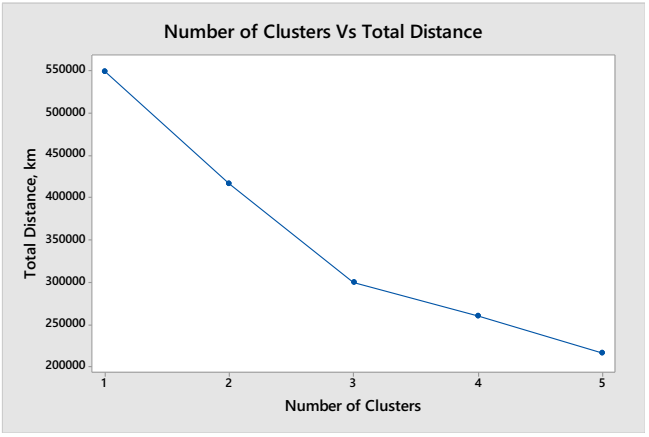

Fig. 4. Preferred Numbers of Clusters.

Five Clusters: B-E-C-A-D.

For three to five clusters, we get the same ranking, B-E-C-A-D. As expected, algorithm B that uses the Weiszfeld's algorithm accounts for the minimum total distance in all the five cases. The performance of the algorithm 'E' that uses Manhattan distance instead of Euclidean distance is also reasonably better. Algorithm 'D' which is the Weighted Weiszfeld'd algorithm reports maximum total distance in 3 of the five cases and is the worst performer as far as the total distances are concerned.

The cluster centres computed by the algorithm 'B' (Weiszfeld's algorithm) are converted to the geographic locations and presented in Table 3. They are graphically plotted in the Google maps and presented in Figs. 2 and 3. Table 3 also shows the total distance reported by the better performer, algorithm 'B' and the percentage difference between the minimum and maximum total distances obtained from different algorithms. The difference is minimum (0.671%) for two clusters and maximum (3.263%) in the case of four clusters.

When the total distance is plotted against the number of clusters (Fig. 4), it is found that k = 3 may be the preferable number of clusters as the slope significantly starts decreasing after k = 3 (Elbow Rule). The gap (difference in the total distance) between the number of clusters 2 and 3 is wide whereas, the gap is less between the number of clusters 3 and 4.

8. Conclusion

This paper proposes a model to find the clusters and their centres from the known geodetic coordinates. 661 Indian districts are clustered and analyzed. If only one cluster is to be estimated, it is the single geometric median and approximately lies at India's centre. The paper estimated up to five clusters and their centres to have a minimum total distance. Both Euclidean and Manhattan distances are used in the computation.

These centres shall help in planning the future and establishing nodal centres for specific activities. If a business is a population-based (a consumer product company, materials' supply points for common customers, consumer goods like toothpaste, consum-

ables), centres are to be estimated based on population data. Similarly, if the distance is important in establishing any nodal centre, data about the distances are to be used. A typical example is the supply of required raw materials like raw materials, water to a manufacturing plant. The number of data points/ districts attached to any cluster varies from cluster to cluster as well as algorithm to algorithm. Up to three clusters, the decrease in the total distance is steep and then it starts decreasing.

Using the 'Elbow Rule', it is concluded that having three clusters will be preferable as establishing more clusters/ facilities will increase the total cost also.

Taking the population percentage share as the weights, clusters are computed in one case, which can be extended for other weights like GDP, literacy rate.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Weber, On the location of industries, Progress in Geography 6 (1) (1982) 120–128.
[2] H.W. Kulin, R.E. Kuenne, An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics, Journal of Regional Science 4 (2) (1962) 21–33.
[3] L.N. Tellier, The Weber problem: solution and interpretation, Geographical Analysis 4 (4) (1972) 215–233.
[4] A. Baskar, Analysing a few trigonometric solutions for the Fermat-Weber facility location triangle problem with and without repulsion and generalizing the solutions, International Journal of Mathematics in Operations Research 10 (2) (2017) 150–166, https://doi.org/10.1504/IJMOR.2017.081922.
[5] E. Weiszfeld, Sur le point pour lequel la somme des distances de n points donnes est minimum, Tohoku Mathematical Journal, First Series 43 (1937) 355–396.
[6] P. Chatzoglou, D. Chatzoudes, Z. Petrakopoulou, E. Polychrou, Plant location factors: a field research, OPSEARCH 55 (3-4) (2018) 749–786.
[7] D.Z. Chen, H. Wang, New algorithms for facility location problems on the real line, Algorithmica 69 (2) (2014) 370–383, https://doi.org/10.1007/s00453-012-9737-0.
[8] Mahdian, M., Pal, M. (2003). Universal facility location. In: ESA 2003, Lecture Notes in Computer Science, 2832, pp. 409-421.
[9] M. Dzator, J. Dzator, An effective heuristic for the P-median problem with application to ambulance location, OPSEARCH 50 (1) (2013) 60–74, https://doi.org/10.1007/s12597-012-0098-x.
[10] Latitude and Longitude. https://www.mapsofindia.com/lat_long/, accessed on Jan. 19, 2019.
[11] C.C. Robusto, The cosine-haversine formula, The American Mathematical Monthly 64 (1) (1957) 38–40.
[12] The MathForum. http://mathforum.org/dr.math/, accessed on Jan. 26, 2019.
[13] Accurate Conversion of Earth-Fixed Earth-Centred Coordinates to Geodetic Coordinates. https://hal.archieves-ouvertes.fr/hal-01704943/document, accessed on Jan. 30, 2019.
[14] National Hurricane Center and Central Pacific Hurricane Center. http://www.nhc.noaa.gov/gccalc.shtml, accessed on Jan. 30, 2019.
[15] L. Cooper, I.N. Katz, The Weber problem revisited, Computers & Mathematics with Applications 7 (3) (1981) 225–234.