

# Improved K-means Clustering based Distribution Planning on a Geographical Network

**Manju Mam**

Manav Rachna International University/EEE Dept., Faridabad, 121004, India  
E-mail: manjumam@yahoo.com

**Leena G**

Manav Rachna International University/EEE Dept., Faridabad, 121004, India  
E-mail: leenag.fet@mriu.edu.in

**N S Saxena**

Management Development Institute, Gurgaon, 122007, India  
E-mail: nssaxena@yahoo.co.in

**Abstract**—This paper presents a distribution planning on a geographical network, using improved K-means clustering algorithm and is compared with the conventional Euclidean distance based K-means clustering algorithm. The distribution planning includes optimal placement of substation, minimization of expansion cost, optimization of network parameters such as network topology, routing of single/multiple feeders, and reduction in network power losses. The improved K-means clustering is an iterative weighting factor based optimization algorithm which locates the substation optimally and improves the voltage drop at each node. For feeder routing shortest path based algorithm is proposed and the modified load flow method is used to calculate the active and reactive power losses in the network. Simulation is performed on 54 nodes based geographical network with load points and the results obtained show significant power loss minimization as compared to the conventional K-means clustering algorithm.

**Index Terms**—Distribution planning, kmeans, geographical network, modified load flow, improved kmeans, clustering, feeder routing.

## I. INTRODUCTION

Automated distribution system expansion planning is an active field of research and attracted attention of researchers in the last decade. Switching to automated system due to various reasons has led to various benefits, like minimization of expansion cost, optimal placement of distribution system, and reduction in system's failures/faults. The planning includes optimal routing of single/multiple feeders, optimize network design parameters such as network topology, branch conductor size and the reserve location for future placement of substations. The significant part of the investment cost is closely related to the size of the substation, length of

feeder, losses in the network, and voltage drop subject to the distance between substation and the loads.

In this paper a new distribution system planning algorithm is developed to meet the following objectives:

1. Optimal placement of substation
2. Optimization of feeder length
3. Power loss minimization and
4. Voltage drop improvement

Improved K-means clustering algorithm is developed for the optimal placement of substation as well as to improve the voltage drop and is compared with the conventional K-means clustering algorithm. A new shortest path based feeder routing algorithm is proposed to optimize the feeder length while modified load flow method proposed by Das et. al. [1] is used to calculate the power losses in the network. The structure of the paper is organized as follows: Section II presents the related work, Conventional K-means clustering algorithm is presented in section III, while the improved K-means clustering algorithm is discussed in the section IV. Section V describes the process to find the optimal location of substation. Modified load flow method is presented in section VI. Section VII deals with the shortest path based single/multiple feeder routing algorithm. Results and discussions are presented and described in section VIII. Section IX contains the conclusion while references are provided at the end.

## II. RELATED WORK

Various approaches have been proposed in the past by researchers and scientists to:

1. Minimize the expansion cost
2. Place the substation optimally and,
3. Protect the system from failures.

In 1991 the method proposed by Ramirez-Rosado and T. Gonen [2] used the concept of pseudo dynamics to expand the power distribution systems. The method uses minimization of non-linear optimization problem to meet the objective. In the same year Nara and team [3] has developed a multi-year expansion planning model for distribution systems. The method uses decomposition algorithm based on branch exchange method in which n-years planning problem is decomposed into one-year planning problem, and one-year results are coordinated through.

In the previous decade, efforts have been made on multistage expansions like cost and complexity analysis when execution is performed at once and when performed in phase wise manner (5/10/20 years). Multistage problem is addressed by Ranjan et. al [4] in 2002 and Vaziri et. al [5] in 2004. A generalized optimization framework is designed by Ranjan to address the issue of substation planning, expansion cost, and optimal selection of branch conductor. The optimization framework incorporates heuristic rules and modified load flow [1], while the method presented by Vaziri used graph theory and numeral analysis technique to solve the complex mathematical non-linear optimization problem.

Many researches have been performed in the current decade to minimize the investment cost while considering the system security. A weighted K-means clustering method is developed by Peng and Liu [6] in 2009. The author's goal is to minimize the investment cost while the system is protected by dividing the load points into small clusters. In 2010 Ouyang et. al [7] presented the concept of integration of distributed generation (DG) for distribution network planning. The author's goal for the integration of DG is to cut the peak load to minimize the cost as well as to improve the efficiency of the network. Low utilization efficiency generally occurred during peak load in conventional distribution network which resulted in high risk and high investment. Later some mathematical models have been proposed by some researchers to solve distribution substation planning such as integer linear programming, mixed integer non-linear programming, and network flow programming.

### III. K-MEANS CLUSTERING ALGORITHM

Clustering technique is required in many different applications, such as data mining and knowledge discovery [8], pattern classification and recognition [9], and data compression and vector quantization [10]. K-means clustering is a simple and well known clustering algorithm and is closely related to number of other clustering and location problems [11]. K-means is more efficient and unsupervised learning algorithm for classifying observed data. Euclidean distances is the main factor in evaluating the distance between different points and are then grouped into clusters. In K-means clustering it assumes that object attributes are from a vector space [6]. The objective is to minimize the squared error function or total intra cluster variance as given in the equation below:

$$C = \sum_{i=1}^N \sum_{x_j \in D_i} (x_j - \mu_i)^2 \quad (1)$$

Where,

N is number of clusters

$D_i$  is the data points in  $i^{\text{th}}$  cluster and,

$\mu_i$  is the mean point of centroid of all data points  $x_j \in D_i$ .

K-means clustering divides the complete data into N clusters either randomly or by using some pre-defined centers. After it calculates the mean of all data to find the centroid of each cluster set. Then the centroids are recalculated using the equation (2) and it repeats until the convergence criteria is met (i.e. centroid does not change).

$$\mu_i = \sum_{x_j \in D_i} \frac{x_j}{|D_i|} \quad (2)$$

### IV. IMPROVED K-MEANS CLUSTERING ALGORITHM

As discussed above that K-means forms the clusters on the basis of distance. Loads at nodes points are also an important factor in distribution network planning and plays a crucial role in minimizing the power losses in the system. In order to prioritize the load points, load weight factors ( $wf_j$ ) are given to all data points. The load weight factor is calculated by the iterative process and taken as 2.5% of the node's load if the node's load is greater than or equal to the mean of all loads. On the other hand, the load weight factor is taken as 2.8% of node's load. The new objective for improved K-means clustering algorithm is given as:

$$C = \sum_{i=1}^N \sum_{x_j \in D_i} wf_j (x_j - \mu_i)^2 \quad (3)$$

The centroid will be calculated from the following equation:

$$\mu_i = \frac{\sum_{x_j \in D_i} wf_j x_j}{\sum_{x_j \in D_i} wf_j} \quad (4)$$

### V. OPTIMAL LOCATION PLACEMENT BASED ON K-MEANS AND IMPROVED K-MEANS

Substation optimal location placement is a highly complex and challenging task. The capacity load, annual load growth as well as distance of load points from substation are an important factor in distribution network planning and play a crucial role in minimizing the power losses in the system. Author [4] considered the power loss minimization objective to find the optimum location of substation while Peng et. al [6] considered the weight factors to model their algorithm. Annual growth factor is an important aspect of future planning, hence annual growth factor, capacity load, distance between load points, and power loss minimization is considered in this paper. The algorithm has been applied on a system having 54 nodes points with geometric as well as load information [4]. In order to get the optimal substation location, the

improved K-means algorithm runs iteratively varying weights. The location which gives minimum power loss is considered as best location. Complete approach is discussed in the algorithm below.

1. Initialize the network data
2. Randomly initialize the load weight factor

$$wf = \text{rand} ();$$

3. Calculate the annual load growth

$$ALG = \text{Load} \times (1 + g)^t$$

4. while  $|Ploss(i) - Ploss(i - 1)| < tolerance$ 
  - i. Calculate location using Improved K-means clustering algorithm
  - ii. Compute the Power loss using load flow method (Section V)
  - iii.  $wf = \text{rand} ();$
5. Store the best location for t
6. Compute mean for optimal location of substation

Where,  $wf$  = load weight factor  
 $\text{rand} ()$  = uniformly distributed random number between 0 and 1  
 $Ploss(i)$  = System power loss at  $i^{\text{th}}$  iteration  
 $tolerance = 0.001$   
 $ALG$  = Annual load growth  
 $g = 7\%$  annual growth rate  
 $t$  = number of years (0, 1, 2, 3, ..., N)

### VI. MODIFIED LOAD FLOW METHOD

Generally, the distribution networks are radial and have high  $\frac{R}{X}$  ratio. Due to high  $\frac{R}{X}$  ratio conventional load flow methods like Newton-Raphson [12-15] and fast decoupled (FD) load flow method gives inefficient result. Das [1] proposed a simple and efficient method for load flow solution of radial distribution network. If the voltage difference at every node in successive iteration is less than 0.0001 p.u. then the algorithm will be considered as converged. The solution methodology of the modified load flow method is as follows:

1. Find the sending node and receiving node of the feeder. Figure (1) represents the single line diagram of the distribution network feeder. Its equivalent sending node and receiving node of feeder along with the branch number is shown in Table 1. Electrical equivalent of figure (1) is shown in figure (2).
2. Calculate current, total real and reactive power loads fed through node 2 using the equation (5) - (14)

$$I(1) = \frac{|V(1)|\angle\delta(1) - |V(2)|\angle\delta(2)}{R(1) + jX(1)} \quad (5)$$

$$P(2) - jQ(2) = V^*(2)I(1) \quad (6)$$

From equation (5) and (6)  $|V(2)|$  is calculated as:

$$|V(2)| = \left\{ [P(2)R(1) + Q(2)X(1) - 0.5|V(1)|^2]^2 - (R^2(1) + X^2(1))(P^2(2) + Q^2(2)) \right\}^{\frac{1}{2}} - (P(2)R(1) + Q(2)X(1) - 0.5|V(1)|^2)^{1/2} \quad (7)$$

Where  $P(2)$  and  $Q(2)$  are total real and reactive power loads fed through node 2.

$P(2)$  = sum of the total real power loads of all the nodes beyond node 2 plus the real power load of node 2 itself plus the sum of the real power losses of all the branches beyond node 2.

$Q(2)$  = sum of the reactive power loads of all the nodes beyond node 2 plus the reactive power load of node 2 itself plus the sum of the reactive power losses of all the branches beyond node 2.

Equation (7) can be written in generalized form as:

$$|V(m2)| = [B(j) - A(j)]^{1/2} \quad (8)$$

Where,

$$A(j) = P(m2) * R(j) + Q(m2) * X(j) - 0.5 * |V(m1)|^2 \quad (9)$$

Table 1. Branch number, sending nodes and receiving nodes of above figure

Branch Number	Sending end node	Receiving end node	Branch Number	Sending end node	Receiving end node
1	1	2	8	6	7
2	2	3	9	6	8
3	3	4	10	3	11
4	4	5	11	11	12
5	2	9	12	12	13
6	9	10	13	4	14
7	2	6	14	4	15

$$B(j) = \{A^2(j) - [R^2(j) + X^2(j)] \times [P^2(m2) + Q^2(m2)]\}^{1/2} \quad (10)$$

$j$  is the branch number,  $m1$  and  $m2$  are sending-end and receiving-end nodes respectively ( $m1 = IS(j)$  and  $m2 = IR(j)$ ).

Real and reactive power losses in branch 1 can be given by:

$$LP(1) = \frac{R(1) * [P^2(2) + Q^2(2)]}{|V(2)|^2} \quad (11)$$

$$LQ(1) = \frac{X(1) * [P^2(2) + Q^2(2)]}{|V(2)|^2} \quad (12)$$

Above equations can also be written in generalized

form as:

$$LP(j) = \frac{R(j) \cdot [P^2(m2) + Q^2(m2)]}{|V(m2)|^2} \quad (13)$$

$$LQ(j) = \frac{X(j) \cdot [P^2(m2) + Q^2(m2)]}{|V(m2)|^2} \quad (14)$$

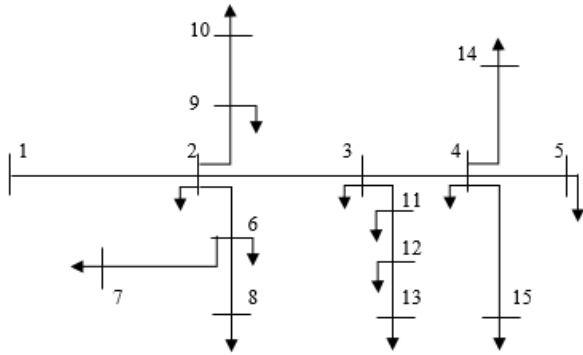


Fig.1. Single line diagram of existing distribution system

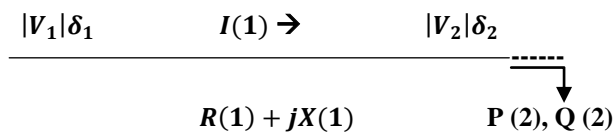


Fig.2. Electrical equivalent of Fig 1

VII. FEEDER ROUTING USING SHORTEST PATH BASED ALGORITHM

The main objective of feeder routing is to minimize the feeder length in order to decrease the cost and to minimize the losses. Resistance (R) and reactance (X) will be required to calculate the power losses in a network. Since R and X is closely dependent on distance as can be seen from the table 2, hence shortest distance based path selection is used to route the feeder throughout the network. Any load point at geometric position X and Y will be connected to its neighbor having minimum distance from the load point. Since the feeder will emerge from the substation so it will find the best neighbor on the basis of distance, and this neighbor will search for another. Once more than 1 node is connected then the node having minimum distance from any one of the connected nodes will be selected for further routing. Suppose we have 10 nodes and a substation as shown in the figure (3) then the substation will be connected to the node having minimum distance from the substation i.e. node 6. Next the node 6 will follow the same process and search its best i.e. node 2. While in the next step both the nodes 6 and 2 will search for the best node and will be connected to node 10 which is best for node 6 and so on. The connection is shown in the figure (4). At the end direction of connections will be stored in a matrix for load flow analysis.

Table 2. Technical information of conductors [18]

Conductor Type	Cross Section (mm <sup>2</sup> )	Resistance ( $\frac{\Omega}{km}$ )	Reactance ( $\frac{\Omega}{km}$ )	Price ( $\frac{INR}{km}$ )	Max Current (Amp.)
Squirrel	12.9	1.376	0.3896	10 <sup>7</sup>	80
Weasel	19.35	0.9108	0.3797	1.2x10 <sup>7</sup>	100
Rabbit	32.26	0.5441	0.3673	1.5x10 <sup>7</sup>	120
Raccon	48.39	0.3657	0.3579	1.7x10 <sup>7</sup>	150

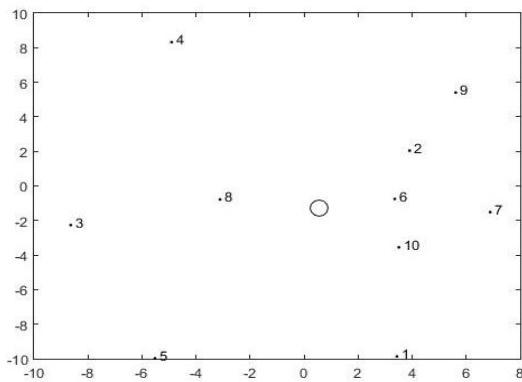


Fig.3. Examples test system having 10 nodes and a substation before connection

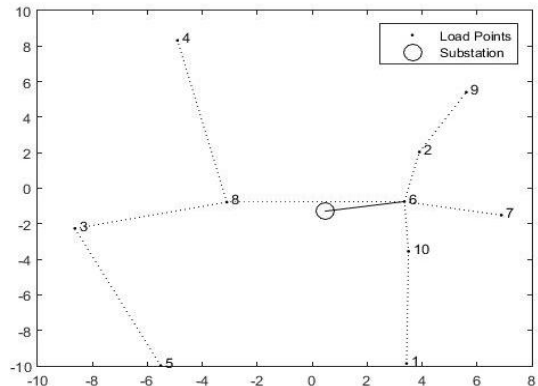


Fig.4. Examples test system having 10 nodes and a substation after connection

VIII. RESULTS AND DISCUSSION

As discussed above the investment cost is directly proportional to length and power losses in the system. Length based shortest path algorithm is developed in order to minimize the power losses as well as the investment cost. In this paper 10 years of planning, single, and multiple feeder routing are considered. 7% of the annual growth factor is considered to model the 10 years of planning and can be achieved by the equation (15)

$$Load \times (1 + g)^t \tag{15}$$

The description of the terms used in the equation (15) are given in the improved K-means section. Power loss is calculated using the modified load flow method. Table 4

contains 54 nodes based geographical network with load points for the simulation and analysis purposes [13].

*Case I. Single feeder routing and substation placement in initial phase:*

In the initial phase only one feeder will emerge from the substation and route throughout the network. The routing will be performed according to shortest path based routing algorithm described in the feeder routing section. Optimal location of substation with single feeder routing with improved K-means and K-means clustering are shown in the figure (5) and figure (6). Total feeder length, active, and reactive power losses are obtained by applying modified load flow method. After modified load flow the results obtained by using both improved K-means and K-means clustering are shown in the Table 3.

Table 3. Substation location, feeder length, active, and reactive power losses in initial phase

Substation location		Total feeder length (in KM)	Active power loss (in kW)	Reactive power loss (in kVAr)	Algorithm
X	Y				
9.0849	11.1226	98.5277	158.7545	44.9497	K-means clustering
9.5874	11.6523	98.5814	11.6503	3.2987	Improved K-means clustering

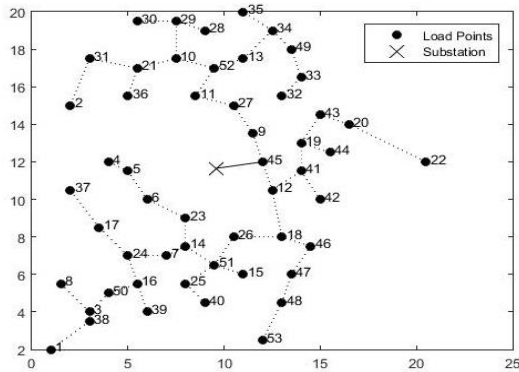


Fig.5. Substation placement and single feeder routing using improved K-means clustering

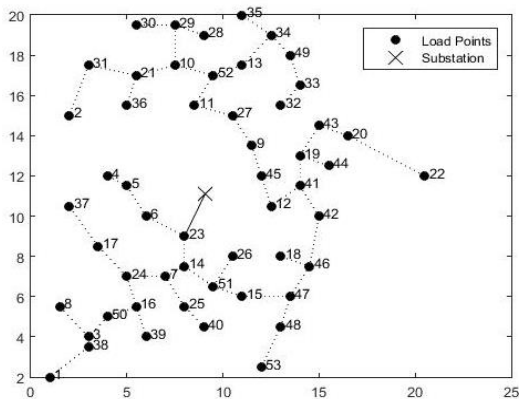


Fig.6. Substation placement and single feeder routing using K-means clustering

Table 4. 54 nodes data points

Node Number	X Coordinate	Y Coordinate	Load (kVA)
1 (Substation)	To be obtained	To be obtained	
2	1	2	25
3	2	15	25
4	3	4	25
5	4	12	50
6	5	11.5	63
7	6	10	63
8	7	7	50
9	1.5	5.5	25
10	11.5	13.5	16
11	7.5	17.5	16
12	8.5	15.5	25
13	12.5	10.5	50
14	11	17.5	63
15	8	7.5	63
16	11	6	25
17	5.5	5.5	16
18	13.5	8.5	16
19	13	8	16
20	14	13	63
21	16.5	14	25
22	5.5	17	25
23	20.5	12	50
24	8	9	100
25	5	7	100
26	8	5.5	100
27	10.5	8	50

28	10.5	15	50
29	9	19	25
30	7.5	19.5	63
31	5.5	19.5	63
32	3	17.5	25
33	13	15.5	50
34	14	16.5	50
35	12.5	19	25
36	11	20	25
37	5	15.5	50
38	2	10.5	50
39	3	3.5	63
40	6	4	25
41	9	4.5	25
42	14	11.5	50
43	15	10	50
44	15	14.5	25
45	15.5	12.5	25
46	12	12	63
47	14.5	7.5	63
48	13.5	6	25
49	13	4.5	16
50	13.5	18	16
51	4	5	25
52	9.5	6.5	16
53	9.5	17	25
54	12	2.5	50

Case 2: Triple feeder routing after 10 years

After 10 years’ substantial amount of load will increase in the network and thus single/double feeders fail to fulfil the demand economically. In order to fulfil the demand economically more than two feeders will be required because total active and reactive power losses will increase up to 496.88% in K-means clustering while 453.31% in improved K-means clustering with single feeder. While in the case of double feeders the amount will increase up to 301.37% in K-means clustering while 275.149% in improved K-means clustering. In this case triple feeders are routed to minimize the active and reactive power losses in the network. The results thus obtained after load flow analysis for both the improved K-means and K-means clustering are shown in the Table 5. Triple feeder routing for both the cases are shown in the figure (7) and figure (8).

Table 5. Substation location, feeder length, active, and reactive power losses after 10 years

Substation location		Total feeder length (in KM)	Active power loss (in kW)	Reactive power loss (in kVAr)	Algorithm
X	Y				
9.0849	11.1226	100.0692	67.5037	19.1130	K-means clustering
9.5874	11.6523	100.2011	55.1342	15.6107	Improved K-means clustering

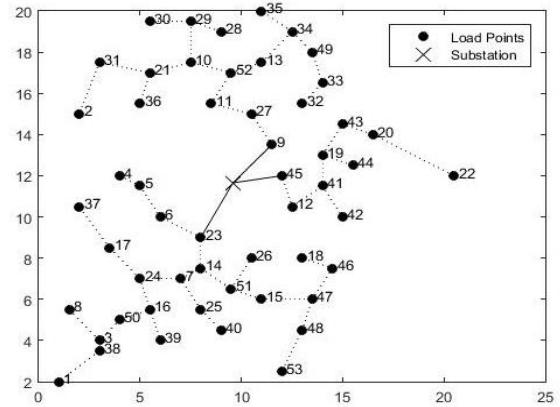


Fig.7. Triple feeder routing using improved K-means clustering

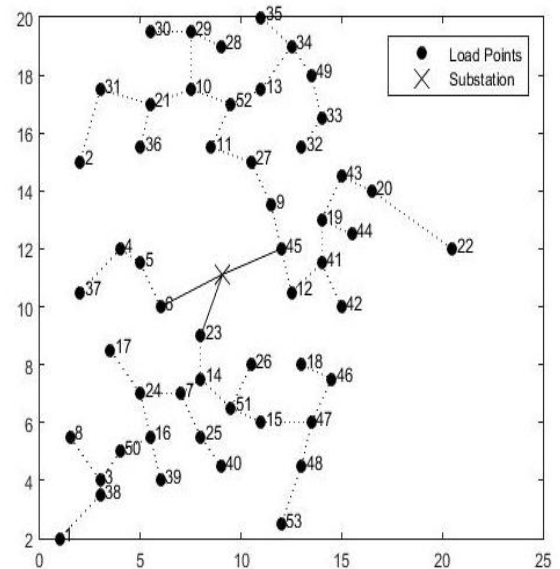


Fig.8. Triple feeder routing using K-means clustering

### IX. CONCLUSION

In this paper a novel improved K-means clustering algorithm is proposed on a geographical network and compared with conventional K-means clustering algorithm as the distribution planning includes the optimal placement of substation, network parameter optimization, power loss minimization, minimization of expansion & investment cost. Distance between substation and load points are the major factor for voltage drop and leads to power losses in the network. A novel Euclidean distance with weighting factor based improved K-means clustering algorithm is proposed in this paper to improve the voltage drop and to minimize the power loss in the network. Single/Multiple feeders are routed throughout the network using shortest distance based algorithm and the modified load flow method is used for the calculation of active and reactive power in the network. 54 nodes based geometric network is used in this paper for results and analysis purposes. The results thus obtained show significant reduction K in power loss as compared to the conventional K-means clustering

algorithm.

#### REFERENCES

- [1] Das D, Kothari D P, Kalam A. Simple and efficient method for load flow solution of radial distribution networks. 1995:17(5):335-346
- [2] Ramirez-Rosado I.J, Gonen T. Pseudo dynamic Planning for Expansion of Power Distribution Systems. IEEE Trans. on Power Systems. 1991:6(1):245-254
- [3] Nara K, Satoh T, Aoki K, Kitagawa M, Yamanaka M. Multi-Year Expansion Planning for Distribution Systems. IEEE Trans. on Power Systems. 1991:6(3):952-958
- [4] Ranjan R, Venkatesh B, Das D. A new algorithm for power distribution system planning. Electric Power System Research. 2002:62:55-65
- [5] Vaziri M, Tomsovic K, Bose A. Numerical Analyses of a Directed Graph Formulation of the Multistage Distribution Expansion Problem. IEEE Trans. on Power Delivery. 2004:19(3):1348-1354
- [6] Peng W, Liu W. A new method for substation planning problem based on weighted K-means. Conference: Advances in Neural Networks. 2009
- [7] Ouyang W, Cheng H, Zhang X, Yao L. Distribution network planning method considering distributed generation for peak cutting. Energy Conversion and Management. 2010:51(12):2394-2401
- [8] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
- [9] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis. New York: John Wiley & Sons, 1973.
- [10] A. Gersho and R.M. Gray, Vector Quantization and Signal Compression. Boston: Kluwer Academic, 1992.
- [11] Kanungo D., Mount D. M., Netanyahu N. S., Piatko C.D., Silverman R., Wu A.Y. An efficient k-Means clustering algorithm: Analysis and implementation. 2002:24(7):881-892
- [12] Tinney W F, Hart C E. Power flow solution by Newton's method. IEEE Trans. on PAS. 1967:PAS-86(11):1449-1460
- [13] B Stott, Alsac O. Fast decoupled load flow. IEEE Trans. on PAS. 1974:PAS-93(3):859-869
- [14] Stagg G W, El-Abiad A H. Computer methods in power system analysis. 1968 Tata McGraw Hill
- [15] Nagrath I.J, Kothari D.P. Modern power system analysis. 1989 2<sup>nd</sup> edition Tata McGraw Hill
- [16] Kalesar B.M, Seifi A.R. Optimal substation placement and feeder routing in distribution system planning using genetic algorithm. Elixir Elec. Engg. 2011:37:3908-3915



**Dr. Leena G.** has B.Tech in Electrical and Electronics Engg. and M.Tech in Control Systems from Kerala University in 1991 and 1995 respectively. She completed her PhD in 2007 from Indian Institute of Technology, Kharagpur, India. She had over 16 years of teaching experience and presently she is a Professor in the Electrical and Electronics Department of Manav Rachna International University, Faridabad, India. Her areas of interest are nonlinear system, decentralized control, sliding mode control, Power Distribution etc.



**Dr. N.S. Saxena.** Professor, Operations Management, MDI, graduated in Electrical Engineering from Madhav Engg. College. He completed Post Graduation in Power Systems and doctoral degree from IIT, Kanpur in 1983. He has over 38 years of work experience in different executive and managerial positions in various Academic Institutions and Central Sector Public Sector Utilities viz. NTPC Ltd. and PGCIL. He retired as Director General of NPTI and his last assignment before joining MDI was Adviser to Hythro Power Corporation Ltd.

**How to cite this paper:** Manju Mam, Leena G, N S Saxena, "Improved K-means Clustering based Distribution Planning on a Geographical Network", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.4, pp.69-75, 2017. DOI: 10.5815/ijisa.2017.04.08

#### Authors' Profiles



**Manju Mam.** Research Scholar, Manav Rachna is B.Tech. in Electronics and Communication Engineering from NIT, Srinagar and M.S in Software Systems from BITS Pilani. She also holds an MBA degree with specialization in HR. She has over 26 years of teaching experience. Her areas of interest are Power Distribution, Geographical Information Systems, IT in Power Industry.