Chapter 4

Clustering of Feeders

4.1 Introduction

Under the deregulation environment, the most needed task for promoting Demand Side Management is to study the continuous load profiles [7]. India being developing country, the electricity usage scenario is totally different than the other developed countries. Looking to the complete irrational and subsidy oriented tariff mechanism, it is necessary for the utility to have load profile study for developing rationalized tariff structure so as to sustain in the fast growing open energy market scenario [95]. Prior to tariff modification, consumer segregation not as per categories as Industrial, Commercial, Agricultural or Residential but based on load profile or consumption pattern is necessary [96]. There are various methods proposed for consumer characterization under load profile study and one of them is clustering technique in which the similar load profiles are clubbed together in form of clusters and these clusters may consist of heterogeneous type of consumers of existing classification. The next step after such classification is to propose various tariff options suitable to the utilization patterns of the consumer sets. Such tariff proposals have to be designed keeping in mind that the utility must not face any financial scarcity in terms of revenue recovery. In short, the new tariff structure is required to be designed with the hypothesis that revenue recovery remains at least same or does not go down, even after offering the new tariff structure [97].

Hence, it has become necessary to segregate the electricity supply feeders based on segregation techniques. Here, two such techniques as Self Organizing Maps (SOM) and k-means clustering are considered. Such segregation will help the utility in forming sub-categories of consumers for offering new tariff mechanism which will be based on consumption patterns.

4.2 Load Profile Study

For network planning and operation, power production planning, demand side management, cost-of-service study, pricing and tariff planning etc proper load profile study at regular interval is necessary [96]. The most needed information for such study is usage of electrical energy at different hours of a day, weeks, seasons and their share to utilities profile. Self Organizing Maps' capacity to classify customers using load profiles has been evaluated in [36]. Consumers chosen are university area, residential zones, medium and small size industries etc. The target set was to group the consumers with similar load profile. Though the study expresses Self Organizing Maps features but looking to the results presented, redefining of space has not been found or it seems that the load profile segregation did not give any new consumer type. Cascade application of Self Organizing Maps based clustering with k-means algorithm has been presented in [37]. Study has been conducted on an Industrial park in Spain. Based on set criterion, total seven clusters were identified for daily load patterns. But the segregation done based on daily, weekly, monthly load patterns into 7×7 clusters using Self Organizing Maps and 5 clusters using k-means algorithm seems to be little complex. Accuracy based assessment of classification tool has been conducted in [38]. The Self Organizing Maps based classification technique and follow-the-leader algorithm are compared. Considering the consumer classes as industry, services and small business activities, follow-the-leader algorithm turns out to be advantageous in easily isolating uncommon load patterns which will be very useful in tariff diversification. The study defines certain indices which are derived from statistics of load diagrams for characterization of electrical customers' behavior [39]. Based on two performance measures, optimal vector is considered from the generic index vectors for giving sharper discrimination properties. The results indicate that some of the consumers were inappropriately billed. A hypothesis of assuming no change in revenue collection from existing tariff and newly offered tariff has been tested for developing a new tariff mechanism. For upcoming market strategies, dynamic tariff possibility has been checked. Results of a self-organizing real-time electricity pricing simulator are presented in [97]. The unit rate has been found as a function of consumption history, grid load and the type of consumer. The simulator works on probability of load reduction or load rescheduling. The cases have been considered as per probability indices ranging from 0 to 1. The optimum criterion is found in the range of 0.6 to 0.4 probability values. The paper presents a stability index for determining the priority rank of clusters [98]. Using three algorithms, initially the clustering has been performed and after testing with respect to defined indices, the most suitable algorithm has been chosen for optimal clustering. Out of k-means, Self Organizing Maps and Fuzzy c-means (FCM) algorithms, k-means turns out to be better in over-all performance. For smaller cluster sizes upto say 6 or 7 clusters, k-means and Self Organizing Maps performance is comparable. For the purpose of a new tariff design, [40] focuses on simulation of system load profile based on integration of various consumer load profiles as per their contribution to system load. Then depending upon consumer class contribution only new capacity cost for each class has been derived. [12] is a study conducted under the various activities of Bureau of Energy Efficiency for implementing Demand Side Management. The residential and commercial establishments in Gujarat were targeted to understand end-use consumption patterns. It has been found out that the actual residential load was terribly higher than the connected load specified and the commercial category had lesser actual load than the declared connected load. The finding pointed out that it was highly necessary for the utilities to have the load corrections for the connected load of the residential category by asking consumers to declare installations. This is how intermediate load profile study enables utility to figure out system changes needed, if any, as per the changing load conditions. Hence, in this study an endeavor has been made to go for segregation of feeders of a chosen distribution sector based on common classification techniques like Self Organization Maps and k-means clustering. It has been tried to analyze the usefulness of results in tariff rationalization process. The area selected is the Dadara Nagar Haveli Power Distribution Corporation Limited (DNHPDCL).

4.3 Clustering Techniques

Clustering is one of the methods of data mining. Various clustering algorithms such as k-means, hierarchical, density based, self-organizing maps etc are mentioned in [99]. Density based clustering involves auto clustering in which number of clusters are automatically formed as the algorithm proceeds. Under the K-means and Self Organizing Maps techniques, number of clusters are predefined and data set is clustered as per the set criterion. Following write-up portrays basic clustering mechanism.

For the group of selected N samples, the complete data-set is represented as a matrix eq.(4.1) [97]:

$$X = \{x_{nh}, n = 1, \ldots, N; h = 1 \ldots, H\} \tag{4.1}$$

Where,

N is total number of samples

H is the time domain for the data as 15 min, 30 min or 1 hour

Basic rule of clustering is segregation as per the similarity between the data vectors / samples which is calculated using say Euclidean distance between the cluster centers [40]. Clustering or segregation of consumers can be done based on availability of load data from the various type of consumers, energy injection voltage level etc. Indian electricity system has feeder level data monitoring and recording. Hence, it is possible to segregate feeders only as per their consumption patterns.

Basic clustering procedure is presented by following steps.

- **Data Selection**: Power consumption data of consumers can be selected as per seasonality, time domain recordings and loading capacity of feeders.

- **Data Cleaning**: Removal of load curves having redundant data, removal of unreasonable load data like holidays, days of network failure etc.

Majority of the feeders are express feeders in this case. Looking to the history of the data collected, it is clear that loading has not changed since long and every feeder represent unique characteristics with no much redundancy.

- **Data pre-processing**: As the clustering is based on the discrete values of consumption over the pattern, data normalization is done within the given range like [-1, 1], [1, 0] depending upon the pattern. It is a method of re-scaling of the data for comparison purpose. min_max, $z - score$, $decimal - scaling$ are few normalization techniques. Here, the data is re-scaled using min_max technique. The normalization is done as per eq.(4.2) in the range of [1, 0] [99].

$$x' = \frac{(x - min_a)\,(new_{max_a} - new_{min_a})}{max_a - min_a} + new_{min_a} \tag{4.2}$$

- **Clustering and Analysis**: Results are obtained by multiple run till the convergence criterion is fulfilled.

  Two algorithms i.e. Self Organizing Maps and K-means, to segregate the data and Silhouette technique to measure effectiveness of the algorithm are described as follows.

4.3.1 Self Organizing Maps

It is Artificial Neural Network technique (ANN) using unsupervised learning method. It transforms high dimensional space to 2-D map [40]. The steps in algorithm are as follows:

1. Randomize the weight vectors of the map nodes

2. Initialize input vector

3. Use the Euclidean distance formula to obtain the similarity between input vectors and weight vector.

4. Track the best matching unit and update the nodes in the neighborhood of it by pulling them closer to the input.

$$\omega(t+1) = w(t) + \phi(t)\alpha(t)[D(t) - \omega(t)] \tag{4.3}$$

5. Go to next iteration and repeat the process from step 2 till the epoch count is over.

Where,

t	current epoch
$\omega(t)$	current weight vector
$D(t)$	input
$\phi(t)$	neighborhood function
$\alpha(t)$	learning restraint due to time

4.3.2 K-Means Algorithm

A highly efficient large data processing technique. As per the set input parameter "K", the set of "N" objects is portioned into "K" clusters. From the input data, initially random "K" objects are selected as cluster centres or means and remaining objects are grouped in the neighborhood as per similarity criterion. New centres are selected and the process is re-iterated until the convergence criterion of squared error is fulfilled [99].

$$E = \sum_{j=1}^{K} \sum_{\rho \epsilon c_j} \mid \rho - S_j \mid^2 \tag{4.4}$$

Where,

E sum of the square error for all the objects in the data set

K number of clusters

ρ point in space representing a given object

S_j mean of cluster c_j

4.4 Measure of Similarity

Silhouette value of a point is the measure of how similar the point is in its own cluster when compared to the points in other clusters. The range of measure is [+1, -1]. Positive higher range indicates well-matched points to its own cluster and poorly matched to the points lying in the neighborhood clusters. Negative value indicates wrong placement. Silhouette value of an i^{th} point is measured as [100]:

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{4.5}$$

For the given range of [+1,-1],

$$S_i = \left\{ \begin{array}{ll} 1 - \frac{a_i}{b_i} & \text{if } a_i \leq b_i \\ 0 & \text{if } a_i = b_i \\ \frac{b_i}{a_i} - 1 & \text{if } a_i \geq b_i \end{array} \right\} \tag{4.6}$$

Where,

a_i average distance from i^{th} point to other points in the same cluster

b_i minimum average distance from the i^{th} point in a different cluster, minimized over clusters

Table 4.1: Feeder Types

Feeder	Type
1	Express feeder at 66kV
2	Express feeder at 11kV
3	Industrial feeders at 11kV
4	Town feeder at 11kV
5	Government office feeder at 11kV

4.5 Result

The feeder level hourly data of five substations of DNHPDCL region have been selected for clustering. Load data of 51 feeders having 24 hour stamping have been used to train the network using both the k-means and Self Organizing Maps algorithms for forming 2 to 7 clusters. Table 4.1 shows the feeder number and corresponding type. Table 4.2 shows feeder segregation in various clusters for k-means algorithm. Table 4.3 shows the mean Silhouette values for both the algorithms. It also presents total feeders having negative Silhouette meaning wrong placement of feeder in corresponding cluster.

Figure.4.1 and 4.2 show the Silhouette plots for the feeders / points falling in clusters formed by k-means and Self Organizing Maps (SOM) algorithms. Silhouette value shows how well the clusters are formed. Figure.4.3 shows the plot of SOM hits representing the hexagonal shape clusters with number of samples falling in each. From the results, it has been seen that mean Silhouettes are closer to unity, in the range of 0.78 to 0.96, for Self Organizing Maps algorithm and having poor values in the range of 0.2 to 0.3 for the k-means algorithm. While focusing on Self Organizing Maps' results, it is clear that wrongly placed feeder cases i.e. case of 3 and 4 clusters, have lesser

Table 4.2: Feeder allocation as per number of clusters

		Number of Clusters					
Feeder	Type	Two	Three	Four	Five	Six	Seven
1	1	1	1	2	2	2	2
2	1	2	2	3	3	4	4
3	1	2	2	3	2	3	3
4	5	2	3	4	5	6	7
5	3	2	3	4	4	5	6
6	3	2	3	4	4	5	6
7	3	2	2	3	3	4	4
8	2	2	3	4	5	6	7
9	2	2	3	4	5	6	7
10	2	2	3	4	5	6	7
11	2	2	2	4	4	5	5
12	2	2	3	4	5	6	7
13	4	2	2	3	3	4	4
14	4	2	2	3	3	4	4
15	4	2	2	3	3	4	4
16	4	2	2	3	3	4	4
17	4	2	2	3	3	4	5
18	2	2	3	4	5	6	7
19	3	2	3	4	4	5	6
20	3	2	3	4	5	6	6
21	3	2	2	4	4	5	5
22	3	2	3	4	4	5	6
23	1	2	2	3	3	4	4
24	1	2	3	4	4	5	6
25	1	2	2	3	3	4	5
26	4	2	3	4	4	5	6
27	2	2	3	4	4	5	6

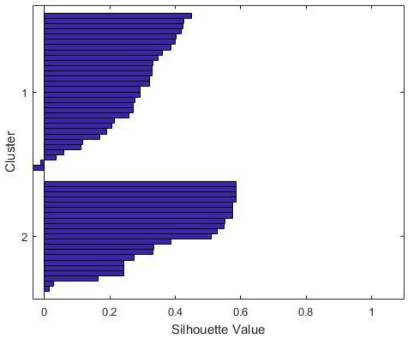
Contd. . . .

Table 4.2 . . . Contd.

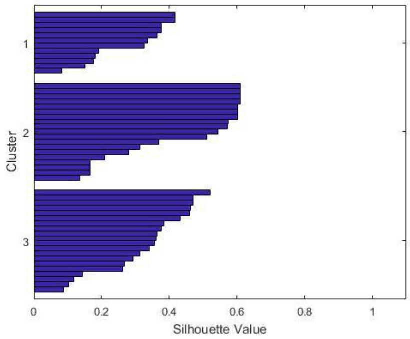
		Number of Clusters					
Feeder	Type	Two	Three	Four	Five	Six	Seven
28	2	2	2	3	3	4	5
29	2	2	3	4	5	6	7
30	3	2	2	3	3	4	4
31	2	2	3	4	5	6	7
32	2	2	3	4	5	6	7
33	3	2	3	4	4	5	6
34	3	2	3	4	5	6	7
35	3	2	2	3	3	4	5
36	2	2	3	4	4	5	6
37	3	2	2	3	3	4	5
38	2	2	2	3	3	4	5
39	2	2	3	4	5	6	7
40	2	2	3	4	4	5	5
41	2	2	3	4	5	6	7
42	1	1	1	1	1	1	1
43	3	2	2	3	2	3	3
44	2	2	3	4	5	6	6
45	3	2	3	4	5	6	7
46	2	2	2	3	4	5	5
47	2	2	3	4	4	5	6
48	2	2	2	3	4	5	5
49	2	2	3	4	5	6	7
50	2	2	3	4	5	6	7
51	3	2	2	3	3	4	5

Table 4.3: Mean Silhouette values for k-means and SOM analysis

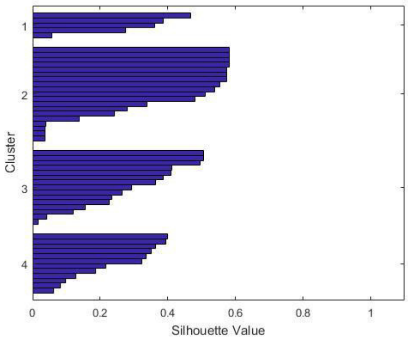
Clusters	2	3	4	5	6	7
k-means Results						
Mean Silhouette	0.3223	0.3575	0.3176	0.3145	0.2875	0.2211
Points giving negative Silhouette value	2	0	0	4	4	7
SOM Results						
Mean Silhouette	0.9606	0.6765	0.6833	0.7872	0.7826	0.7894
Points giving negative Silhouette value	0	3	3	0	1	0



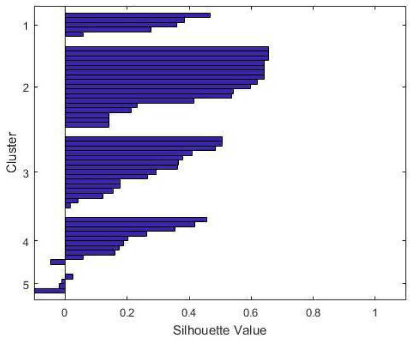
(a) Silhouette plot 2-clusters



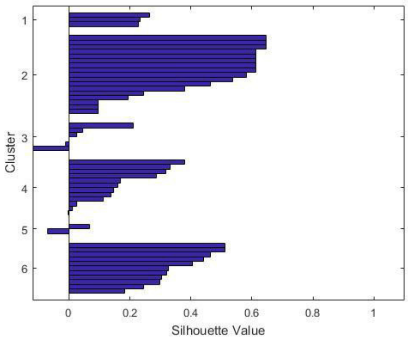
(b) Silhouette plot 3-clusters



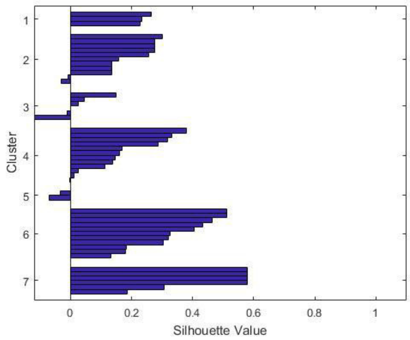
(c) Silhouette plot 4-clusters



(d) Silhouette plot 5-clusters

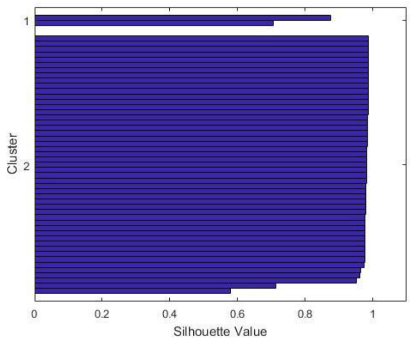


(e) Silhouette plot 6-clusters

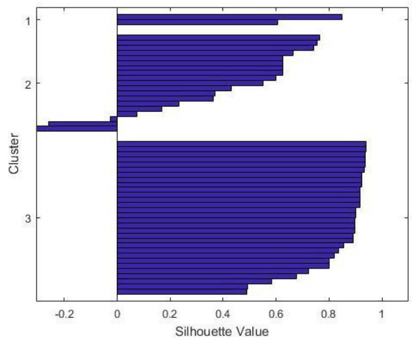


(f) Silhouette plot 7-clusters

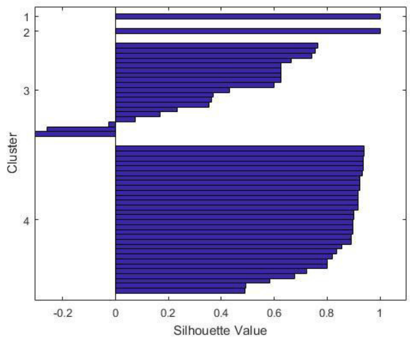
Figure 4.1: Silhouette Plots For k-means Clustering



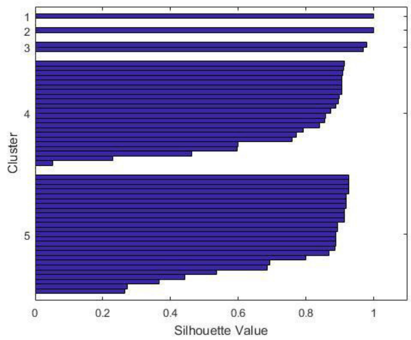
(a) Silhouette plot 2-clusters



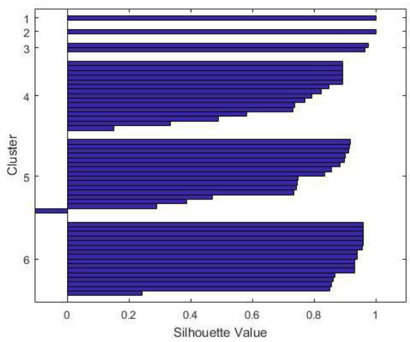
(b) Silhouette plot 3-clusters



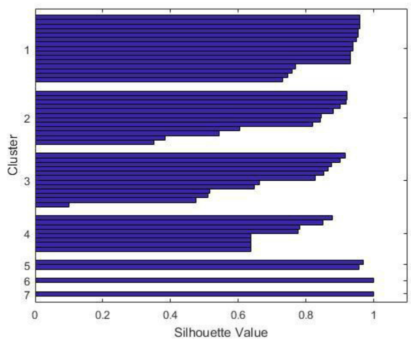
(c) Silhouette plot 4-clusters



(d) Silhouette plot 5-clusters

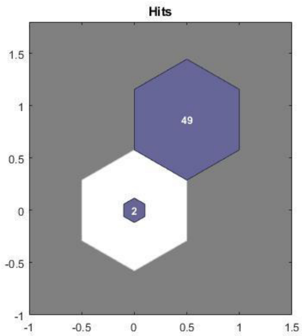


(e) Silhouette plot 6-clusters

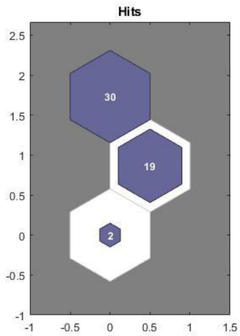


(f) Silhouette plot 7-clusters

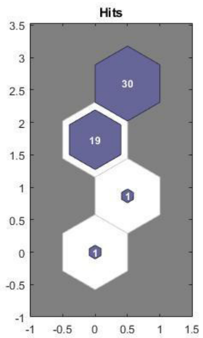
Figure 4.2: Silhouette Plots For SOM Clustering



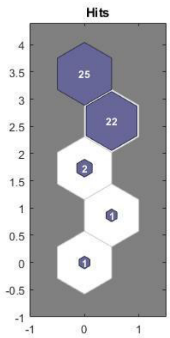
(a) SOM plot 2-clusters



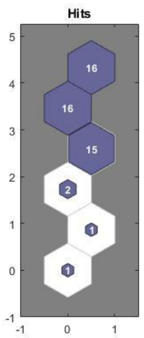
(b) SOM plot 3-clusters



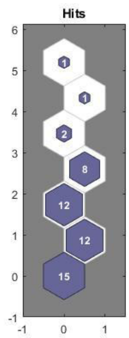
(c) SOM plot 4-clusters



(d) SOM plot 5-clusters



(e) SOM plot 6-clusters



(f) SOM plot 7-clusters

Figure 4.3: SOM plots showing number of feeders in each cluster

Silhouette compared to other cases. As the number of wrongly placed feeder increases, the Silhouette number reduces. Better the Silhouette, algorithm performance is better.

4.6 Conclusion

Though higher silhouette values have been obtained for the six cases of Self Organizing Maps algorithm, not more than 2 cluster formation is acceptable. This is due to presence of misplaced feeders for 3 and 4 cluster formation case and beyond this number, it is meaningless to form higher clusters i.e. consumer segregation, due to limited number in consumer types. Though the k-means results have no misplaced feeders for 3 and 4 cluster formation, the Silhouette is quiet poor. Hence, the process of clustering does not give significant results with the available data for majority of industrial feeders. It is needed to perform the same process with more consumer types having comparable loading on system in an urbanized area.