2. CONSUMERS LOAD PROFILE WITH DATA MINING

2.1 Basic Load Profile Analysis

A load profile is described as "the pattern of electricity load demand of a consumer or a group of consumers over a given period", in which time interval could be day, week, month, or year. The idea behind is to use a time slots as the effective tool for system planning, tariff rate formulation, devising marketing strategies and load management. The analysis has been categorised as power utility consumers which depends on consumer's electricity consumption behaviour [12, 100].

In many countries electricity consumption profiles recognised as a substitute, price-signal approach to the mean time metered solution which is inappropriate and costly for lower & medium voltage, domestic & trading consumers. Power consumers load profiles helps power utilities to determine the power cost, it helps Power utilities to improve efficiency, planning & trading approach.

2.2 Load Profile Approaches

In this case the Load profiles are categorised into two models which are area model and category model which is based on Different regions & on equivalent power consumption categories. The limitation of the geographic region based on area model is that all the power consumers have same load patterns because they are supplied from the same power substations. While the Category model has its own limitations, which is related to dissimilarities in the rest curve and power consumption pattern that's why based on the rest curve & particular power consumption profile found new model device put forward for load profile based settlement services. To get the benefits from both the above given models the third model was designed which was more advanced and efficient. Then too the modelling power consumers load profile the category model would be applied reason being it was strongly believed that the mode would have the greater cognizance into the power consumers demand model.

In some of the countries the innovative technology has come up with lot of modern approaches to power consumption profiling. The two groups were profound in determining particular power consumption profile by different approaches in which the first set of model was obtained from the complete power consumption survey & was developed particular load profiles Different pattern recognition techniques have been used as tool to determine the profile of power consumption to acquire particular power consumption profile based on types of power consumptions techniques noted in the second group.

The first group has its own limitation with the measurements time consumption & the other group has the constraint that the actual method to make the consumer attributes is costly & lengthy. In the previous case because of the calculation was requiring more time & the alternative analysis & it's features of the consumer groups was analysed, the time required for determination is quite high.

To develop power consumption profile various approaches used some of them are very costly & lengthy in second approach better than previous one which is still not apparent & also there is no particular method to understand the optimal solution many techniques have been used to get the required output that's why it is necessarily required to get to understand the study which supports the techniques for identifying, detecting & forecasting irregularities in power consumption behaviour, here we are using MGVCL Gujarat power consumers as the relative sample.

2.3 Load Profiles and Data Mining Techniques

Clustering techniques, known as unsupervised learning, provide starting procedure for solution in examining data analysis and add up with pattern recognition methods. The general purpose of cluster analysis is to identify frequent patterns or to add up related cases through a process of incorporating a set of objects into clusters. In first cluster object should be similar comparing the object in other cluster should be dissimilar is intended outcome. This procedure is not only useful tool for finding the distribution of patterns and interesting correlations among data attributes, but it also acts as an outlier detection tool to identify and detect objects that deviate from normal patterns [70].

In many real-world applications Clustering techniques have been widely used including document clustering, gene expression micro-array data analysis, and image segmentation. Also, they have commonly been used in power utility applications, particularly in load profiles studies, to group analogous load profiles for various purposes. These have included developing better marketing strategies, properly designing tariff structures, and allocating typical load profiles (TLPs) to form groups of eligible consumers. Up till now, no study uses a classification process to group load patterns for each individual consumer based on behavioural similarity as a means of establishing normal and abnormal load patterns for identification and detection benchmarking purposes.

Clustering techniques broadly come into two methods, partitional and hierarchical, with K-means and single linkage widely used in the domains of partitional and hierarchical clustering, respectively. However it is expanded the classification into five main clustering categories, including partitioning methods, hierarchical methods as posited by Chameleon, density-based methods, grid-based methods, and model-based methods such as COBWEB. By contrast, Jain et al. presented a clustering overview from a statistical pattern recognition perspective, including consideration of 1) Agglomerative vs. Divisive, 2) Monothetic vs. Polythetic,3) Hard vs. Fuzzy, 4) Deterministic vs. Stochastic, and 5) Incremental vs. Non-incremental. Among all the clustering categories, there are some advantages and disadvantages depending of each clustering technique on the problem to be addressed and the assumptions made. The mostly accepted clustering techniques used to classify electricity consumers based on their load profiles are set out below [27].

2.3.1 **Fuzzy Clustering** – It is evident that FCM can handle outliers efficiently. This is important in the present research as outliers are to be defined as those patterns that are different from other patterns that are designated as normal. In addition, FCM has been used for Magnetic Resonance Imaging (MRI) segmentation. FCM provides for maximal membership values with a certain degree specified, where it is flexible and reasonable to reject those patterns with low maximal membership

degrees that fall below a certain specified degree. First limitation of this technique is to specify number of clusters in advance

2.3.2 **Hierarchical Clustering–** Unlike Fuzzy Clustering, the main advantage of this technique is that number of cluster is not to be specified previously. Gerbec [101-104] and Chicco [105-107] use hierarchical clustering to group consumers based on their behavioural similarities and apply it for comparisons purposes.

2.3.3 **K-Means Clustering** – The K-means [71] is an iterative scheme that is well known for its efficiency and ability in the clustering of large datasets. For understanding load profile classes in electricity markets, it has been used widely. In K-means application arises two major issues applications, namely 1) require to determine the K cluster in advance as an input, and 2) its use of the alternating minimization method to solve non-convex optimization problems in finding cluster solutions. The improved K-means has been used for exploring local protein sequence motifs so as to represent common structural properties.

2.3.4 **Self-Organizing Maps Clustering –** A self-organizing map (SOM) is an unsupervised neural network learning algorithm. It has proven to be an excellent tool in the exploratory phase of data mining and it has been successfully used for analysis and organization of large data files in various fields, including consumer classification in electricity markets. SOM clustering has been used to form different consumer classes that allow better marketing strategies to be devised. When compared to hierarchical clustering it is a fast and convenient method and has well-known visualization properties.

All-inclusive comparisons among all these clustering techniques have been presented. None of the clustering technique that is universally applicable in exposing the variety of structures present in multidimensional data sets. Consequently, there is a need to conduct a pertinent empirical study to evaluate their performance and suitability in this context.

Arbitrariness of the numbers is main problem in clustering that arises because the data can admit clusters in different sizes and shapes in an n-dimensional data space. The common solution for this is to form a view of the data itself. It is not possible to view clusters in highly dimensional data, in real-world applications. Most load profile studies have adopted the criteria set by Allera: 1) number of clusters large enough to cover the whole sample population, but small enough to be practical, 2) homogeneous representation of the particular load patterns, and 3) significant differences between specified clusters.

Several clustering validity measures have been used to assess the adequacy of the procedure, including Mean Index Adequacy (MIA, Similarity Matrix Indicator (SMI), Clustering Dispersion Indicator (CDI), Davies-Bouldin Index (DBI), Modified Dunn Index (MDI), and Scatter Index (SI). These indicator measures are used to compare the results obtained from different clustering techniques and they tackle the problem by finding an optimal number of clusters. In such an optimum, the number of clusters should be as low as possible, but not so low that it can cause an excessively large dispersion of clusters. A common characteristic of these indicators is that lower values of the indicator correspond to better clustering validity. However, the clustering results are only significant if the number of consumer classes formed by various algorithms is the same. The clustering algorithm that produces the smaller indicator values then prevails over the others in terms of performance [68].

From the previous studies cited above, it is readily apparent that load profile investigations have been attempted for a various reasons. None of the case study of the consumer behaviour changes are considerable for electrical supply entities. These changes may due to many factors and they may signal non-technical losses that flow from irregularities traceable to consumer's actions. Most of the studies mentioned used clustering techniques to group the consumers in their samples accordingly, but none has applied classification techniques. The challenge to be taken up in this study is to recognize that consumer load behaviour can result indifferent consumption patterns at different times. Therefore, accurate load studies that relate to the power distribution system require considerable care. In particular, they must be handled in a manner that ensures that the results obtained do not inadvertently insult legitimate consumers whose behaviour variations do not involve any acts that contribute to an observed non-technical loss pattern.

2.4 Electricity Losses

In general, affected electricity utilities of power losses are categories into two categories: nontechnical losses (referred as an unmetered power losses) [15] and technical losses. Power losses are defined as the difference between powers supplies quantities recorded as sold to consumers and power supply quantities delivered.

The power registered as consumed should equal the electrical power generated. In real scenario, the condition is not similar because losses occur as an integral outcome of distribution losses and energy transmission. Davidson [108] developed these power losses in terms of the as following equations.

Power Losses is defined as follows:

$$E_{Loss} = (E_{Delivered} - E_{Sold})$$

Technical losses arise because of the physical climacteric of power generation, T & D and turbine efficiency covers degrees in generation, transformer, together with substation and line related losses. These involved resistive losses in the primary line (IR), resistive losses and the distribution transformer losses (windings and resistive losses in core losses) in the secondary line, losses in kWh metering and resistive losses in service drops.

Credit loss because of technical losses

$$C_{Loss} = (U_{Electricity\ Cost} \times E_{Loss}) + M_{Maintenance\ Cost}$$

In general we can categories in two types of technical losses: i) the no-load losses are independent of the load served by the system. The major no-load losses are because of transformer core losses leads to excitation current flows and ii) load losses consisting of the I^2R and I^2X losses in the series impedances of the various system elements, although when the system is unloaded, these load losses are obviously non-existent. Unmetered Power losses corresponds to power theft in one form or another which are related to the consumer management process and can include a number of means of consciously defrauding the utility concerned [19].

Unmetered Power Losses (UPL)

$$C_{Unmetered\ Power\ Loss} = (C_{Loss} - C_{Technical\ Loss})$$

Above typical important of fraud perpetration are as shown in Table 2.1

Table 2.1-Unmetered power losses type basis of components identified

Components	Power Utilities	Electricity Consumers
Meter	Inadequacies and inaccuracies of meter reading.	Unauthorised line tapping and diversion.
	Losses due to faulty meters and equipment.	Stealing by bypassing the meter or otherwise making illegal connections.
	Inadequate or faulty metering.	Tampering with meters to ensure the meter recorded a lower consumption reading
	Loss/damage of equipment/hardware, e.g.: protective equipment, meters, cables/conductors, and switchgear.	Faulty meters not reported.
Bills	Inaccurate consumer electricity billing.	Non-payment of electricity bills.
	Inefficiency of business and technology management systems.	Adapting billing irregularities with the help of internal employees.
	Arrangement of billing irregularities with the help of internal employees.	Manipulating readings by bribing meter readers.
	Poor revenue collection techniques.	Inaccurate estimation of non-metered supplies, e.g.: public lighting, agricultural consumption, rail traction.
	Making out lower bills, adjusting the decimal point position on bills.	Avoiding unpaid bills.

In all such cases, electricity consumers intentionally ignore paying their bills or are involved in pilferage, theft, and unauthorised use. The detection and prediction of Unmetered Power Losses activities on the distribution level is the aim of the current study is to focus where deviations in consumer behaviour are found to exist.

2.5 Fraud Detection

In this chapter, "According to user population in order off fraud detection [17] involves monitoring the behaviour of to estimate, detect or avoid abnormal behaviour". Literature examined based on, an analysis of fraud detection and identification techniques is listed below, with professions categorised into two types as follows.

- Electricity professions.
- Type of profession like telecommunications, risk management, credit card provision and insurance.

For the first types, very large numbers of data mining research review on fraud detection and identification in electricity professions are analysed, artificial neural network [16], decision trees, accommodating rough set, statistical-based outlier detection, multiple classifiers and

wavelet-based feature extraction. Several of this studies used data mining algorithm by directly determine them to consumer databases as inputs. A combination of multiple classifiers and wavelet techniques have been applied to identify fraud in a power distribution network. Maximum accuracy is obtained with wavelet technique over easy methods because of its capacities in multi and localization resolution study. As another option, rough sets & decision tree were used respectively for the classification of power consumers. Study also conduct using statistical based outlier mining and the artificial neural network [57], where both studies review a method employing a common framework that had consumer databases as its input data source.

For the other types, many more data mining algorithms used in the other types of professions referred above as, insurance, telecommunications, credit card provision and risk management. In every study, data mining has been determined as a tool that capable of the prediction and detection of fraud. It is noted that in general the credit card company's case study have been that they using the neural network as a tool to detect fraud, with the Security and Exchange Commission and telecommunication businesses applying a similar algorithms. From the above these applications applied data mining techniques to detection of fraud directly from their consumer databases.

2.6 Fraud Detection Techniques

In this review, according to two surveys compares of fraud detection techniques [11] have been analysed on. Hodge [118, 119] proposed the three fundamental types to the problem of specifies detection as mentioned below:

1) Unsupervised

Determine the outliers with no prior knowledge of the data using unsupervised clustering; unsupervised outlier detection has been applied in the context of time series financial data.

2) Supervised

Model both normality and abnormality using supervised classification with pre-labelled data.

3) Semi-supervised

Model only normality or, in a few cases model abnormality, using semi-supervised recognition or detection. These approaches encompass distance-based, set-based, density-based, depth-based, model-based and graph-based algorithms.

4) Statistical-based Outlier Detection

Statistical-based outlier detection identifies outliers using a discordancy test that assumes a distribution or probability model for given datasets. Although it is the simplest approach to outlier detection, its major drawback is that most tests are only for single attributes or one-dimensional samples. This limitation renders it unsuitable for use in data mining problems as most current databases in such applications are multidimensional.

5) Distance-based Outlier Detection

Distance-based outlier detection was introduced to counter the main limitations of statistical methods. Several distance-based algorithms have been developed, including the index-based algorithm, the nested-loop algorithm, and the cell-based algorithm. Two parallel algorithms were proposed that encompass distance-based out lie and density-based outlier detection. The former was based on nested loops, along with randomization and the use of a pruning rule, while the latter required only a parameter comprised of the number of the nearest neighbours used in defining the local neighbourhood of the instance.

6) Density-based Outlier Detection

A density-based approach to mining outliers over datasets with different densities and arbitrary shapes was proposed and the results showed that this method has significant speed improvements with comparable accuracy over the current state of the art density-based outlier detection approaches.

7) Clustering-based Outlier Detection

An efficient cluster-based outlier detection method using a vertical data model was proposed and the results showed that this method facilitated the performance of analyses at five times the speed when compared to other contemporary clustering-based outlier detection approaches.

8) Deviation-based Outlier Detection

Deviation-based outlier detection is totally different from the other two immediately above in that it identifies outliers by examining the main characteristics of objects in the group. Two techniques have been reported as used in this case, namely Sequential Exception Technique and OLAP Data Cube Technique. The techniques to detect outliers can be divided into two categories as set out below.

9) Statistical techniques

An informal box plot has been used to detect univariate outliers directly in areas of medical science. Even though it appears to be a simple method, the accuracy of the classification involved was effectively increased after the outlier detection was completed.

10) Neural network techniques

A review of novelty detection based on the use of neural networks has been reviewed. Various applications have focused on neural networks as their approach to detecting fraud or outliers, including fraudulent financial activities exposed by the Securities and Exchange Commission, communication network fraud, fraudulent credit card operations, and cases of management fraud.

2.7 Conclusion

Consumers' load curves in the case of electricity companies vary according to their consumption patterns, with consumers categorized into three main groups as residential, commercial, and industrial. Having knowledge about these consumers is especially important in deregulated markets if any individual company is to be competitive and stay ahead of its rivals. One way of gaining knowledge about electricity consumers is by studying their load consumption behaviour. This can be done through the interval metering.

As an integral part of meeting their objectives to maximize profit and minimize operation costs, power utilities should cope with the common problem of losses that occur within their operations. As noted here, such losses are aptly considered as Unmetered Power Losses and technical losses. The necessity of minimising Unmetered Power Losses is critical as these losses means that added costs are always passed on to electricity consumers. Beyond all the present solutions available, including the SCADA system and field investigations, the present proposal is to use consumer behaviour changes manifested in load consumption variations as a means of highlighting abnormalities that may be indicative of Unmetered Power Losses activities. For this approached, a fraud analysis system similarly to those implemented by other businesses, including credit card provision, bank loan applications, and writing insurance, is approached for implementation by energy utilities. This proposed technique is implemented and tested using data collected from commercial consumers of MGVCL Gujarat.

This chapter we discuss losses in power utilities, including technical losses and those due to Unmetered Power Losses activity, and relating to the impact of Unmetered Power Losses activity from a financial perspective and an economic. Finally, this chapter provided an overview of MGVCL as the largest power utility in Gujarat and outlines its needs with respect to implementing solutions to minimise Unmetered Power Losses activity.

The following chapter will focus on the data mining techniques that will be used in the proposed framework of analysis as the means of specifically identifying, detecting, and predicting those irregularities and abnormalities in consumer behaviour that can be linked to Unmetered Power Losses. These techniques will include feature selection, classification techniques, and prediction techniques.