IETE Journal of Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tijr20

Exploratory Data Analytics and PCA-Based Dimensionality Reduction for Improvement in Smart Meter Data Clustering

Gulezar Shamim & Mohd Rihan

To cite this article: Gulezar Shamim & Mohd Rihan (2023): Exploratory Data Analytics and PCA-Based Dimensionality Reduction for Improvement in Smart Meter Data Clustering, IETE Journal of Research, DOI: 10.1080/03772063.2023.2218317

To link to this article: https://doi.org/10.1080/03772063.2023.2218317

Published online: 31 May 2023.

Submit your article to this journal

Article views: 23

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tijr20

IETE JOURNAL OF RESEARCH
https://doi.org/10.1080/03772063.2023.2218317

Exploratory Data Analytics and PCA-Based Dimensionality Reduction for Improvement in Smart Meter Data Clustering

Gulezar Shamim and Mohd Rihan

Department of Electrical Engineering, ZHCET, AMU, Aligarh, India

ABSTRACT

The smart meter sends the meter readings to the utilities at desired frequency allowing better visibility of consumer electricity consumption behaviour by providing more data points for in-depth analysis and for generating insights using advanced data analytics and data science techniques. The granulated data helps utilities in designing schemes for audience suitable for demand response management to shift the peak hour demand to off-peak hours. In this paper, a method is proposed for load profile segmentation which can be used by utilities for identifying the characteristics of different users and targeting those whose demand curve can be flattened during peak hours with various demand response management schemes. Firstly, exploratory data analysis is done on the cleaned dataset to find the optimal epoch size, understand the distribution of data in each epoch, and use it for dimensionality reduction. For reducing the clustering computation time, dimensionality reduction is done by around 64% using Principal Component Analysis. The first six principal components are identified as carrying maximum variance using the cumulative variance technique in each epoch. Unsupervised Machine Learning based k-means clustering technique is applied to these principal components. The optimal value of k is evaluated using the WCSS technique where $k = 5$ and $k = 3$ for residential and SME users respectively is found. The average silhouette coefficient for residential users is 0.48 and for SME users is 0.51. Hence, well-separated clusters are formed with minimum intra-cluster distance using PCA for dimensionality reduction which is used for load profile segmentation and Post Clustering Analysis.

KEYWORDS

Cumulative variance; dimensionality reduction; elbow method; exploratory data analysis; *k*-means clustering; partitional clustering; post clustering analysis; principal component analysis; silhouette coefficient; unsupervised clustering; within cluster sum of square (WCSS)

1. INTRODUCTION

The advancements in electricity infrastructure and efficiency play an important role in the overall economic development of a country. “Smart Grid” is a potential technology for meeting the rising demand for a reliable, secure, greener, and more efficient supply of electricity. Smart Grid helps electricity systems to become predictive, communicative, and controllable. It enables “two-way communication” between utility and consumers via integrated sensing and communication devices such as smart meters [1,2]. Smart Grid has improved functionalities such as self-healing, automatic demand response management, reduced peak demands, improved security and reliability, and accommodation of various generation and storage units including renewables [1].

For Demand Response Management and reducing peak demands, it is essential to understand the electricity consumption behavior [3]. Smart meters in Smart Grid provide the electricity consumption data at a desired frequency and hence helps in revealing more insights about the behavioral characteristics of consumers. It helps utilities to introduce different pricing schemes based on the time and day of use and overall demand for electricity [4]. One such scheme is dynamic pricing where consumers can shift their usage and take advantage of low electricity tariffs. This helps in gradually flattening the load curve during peak hours and reducing overall peak demands [3]. The smart meter provides consumption data at the desired frequency with multiple data points in an hour which makes the size of data for analysis very big. Total samples in a data set are found by multiplying the number of samples in a day, the number of days for analysis, and the total number of users. This data set contains a large number of samples and hence should be reduced such that the key information is not lost. Hence, one major challenge is to extract useful information from this data and generate key insights. For this, data analytics and machine learning (ML) are used. ML-based smart meter data clustering helps to put consumers showing similar consumption characteristics in one group which can be further utilized by utilities for demand response management. ML-based clustering techniques are usually unsupervised learning methods in

© 2023 IETE

which inferences are drawn from the input data without labeled responses and are hence useful in smart metering. It finds meaningful structure in the data based on similarity and dissimilarity between different data points [5,6]. Identifying the most suitable method for clustering and reducing the dataset in a way that the useful information is not lost is one of the major challenges faced in this process [7].

In literature, various techniques are proposed for analyzing the smart meter data and clustering the load profiles based upon the variation in consumption pattern. Zafar *et al.* [8] proposed a stochastic modeling approach for load profiling of smart meter data. They have used extended k-means for clustering the raw smart meter data. Lee *et al.* [9] proposed a two-stage clustering technique for the segmentation of residential load profiles where in the first stage the demand characteristics like total consumption, peak demand time, and the difference between maximum and minimum demand are used as an input to k-means clustering model and 6 optimal clusters are found. In the second stage, normalized electricity consumption data is fed to the *k*-means clustering model and clustering is done separately for each cluster identified in stage 1. The only disadvantage of using 2 stage clustering is that number of clusters increases and sometimes very few profiles are assigned to one cluster. Alexandar *et al.* [10] have transformed the raw electricity dataset by applying traditional time series analysis methods like wavelet transformation and autocorrelation analysis before giving it as input to the k-means clustering model. In [5] several methods are compared for the clustering of electricity data like Fuzzy c-means, SOM, and Hierarchical. In [11], three hierarchical clustering methods are proposed for capturing the characteristics of the time series-based smart meter data. The input to these clustering models is a set of features extracted to measure dissimilarity using quantile auto covariances, simple autocorrelations and partial autocorrelations. In [12] the wavelet coefficients extracted from hourly smart meter measurements of commercial and industrial users are given as input to the k-means clustering model for identification of distinct load profiles. In [13] a spatiotemporal visual analysis approach is proposed to identify the energy demand shift patterns across various geographical locations. To reduce time series dimensionality, t-SNE is used. For clustering the smart meter data, the convex clustering technique is proposed in [14]. In [15] for deeper insights into household characteristics, features in multiple domains are extracted using discrete wavelet transform. For feature selection, a random forest classifier is used, and the output its output is given to the SVM.

Although in literature various techniques are proposed for clustering the load profiles based upon electricity consumption of different users, they are either focused on clustering of raw smart meter data which increases the overall computational cost, clustering of aggregated load profiles or the sample set on which the model is applied is small and hence scalability of the technique used cannot be proven. In this work, the aim is to identify a method that can be used by utilities for demand response management by targeting the suitable users identified by load profile segmentation in a way that the computational cost is less and the clustering model performance is good. For this, the contributions of this paper can be summarized as follows:

1. Firstly, after the clean-up of the dataset, analysis of electricity consumption data distribution is done using exploratory data analysis to identify the optimal size of epoch for feature extraction and dimensionality reduction. EDA for identifying the epoch size is not reported in the literature although it helps in identifying the optimal epoch size for dimensionality reduction and for understanding the distribution of data in each epoch of different users.
2. Feature set is selected such that it extracts maximum information in each epoch from the raw dataset. Principal Component Analysis is used for feature extraction from smart meter data so that an optimal number of features are extracted in each epoch which reduces the dimensionality of the dataset and increases the reliability of the model. This technique of feature extraction is not used in the literature and it leads to a considerable reduction in the data size while giving better clustering results as compared to the work done earlier in the literature.
3. The unsupervised ML-based clustering model is selected so that it reduces the computational cost while increasing the scalability, accuracy, and efficiency of the technique used for load profile segmentation.

An unsupervised ML clustering technique used in this work clusters the users based on their consumption pattern whose results can be used by utilities for demand response management and tariff design. The clustering results of this model will help the utilities in setting different pricing rules (ToU & DoU).

This paper is organized as follows: Section 2 introduces the approach used for identifying the optimal epoch size and understanding the Time of Use and Day of Use criteria of electricity consumption. It also gives the basic concept of the method used for feature construction and the clustering technique used for grouping similar load profiles. Section 3 of this paper describes the data set used and reports the EDA and PCA outputs. Section 4 shows the approach used in this work for identifying the appropriate number of clusters and further shows the clustering results on the given data. Section 5 concludes the paper.

2. METHODOLOGY

2.1 Exploratory Data Analysis (EDA)

EDA is the process of performing an initial investigation on the data to identify the patterns, test the hypothesis and validate the assumptions using statistical summary and graphical representations. It is an important step to understand the data and gather many insights [15]. For understanding the distribution of energy consumption data of each user in each epoch box plot is used and to identify the correlation in the consumption pattern on different days in each epoch, heat map is used.

A box plot displays the distribution of data using five summary points – “minimum”, “first quartile” (Q1), “median”, “third quartile” (Q3), and “maximum” as shown in Figure 1. Here, “minimum” represents the minimum electricity consumption of a user in an epoch, “Q1” represents the value below which 25% of datapoints fall in an epoch, “Median” represents the value below which 50% of datapoints fall in an epoch, “Q3” represents the value below which 75% of datapoints fall in an epoch and “maximum” gives the maximum energy consumption in the respective epoch. Box Plot also helps in identifying outliers present in the data, finding if data is symmetrical, identifying how tightly packed the data is, and determining the skewness of the data. In this work, the box plot helps in identifying the distribution of electricity consumption data in each epoch, comparing the distribution of different types of users, and finding the Time of Use effect on the distribution pattern of users in a day [15].

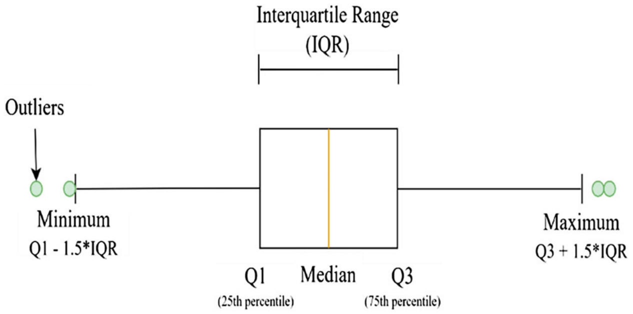


Figure 1: Different parts of the box plot

To understand the effect of Day of Use on consumption patterns in each epoch, the correlation between epochs on different days of a week is found. To find the correlation between the consumption of electricity by a user in respective epochs of two days the following equation is used –

$$r_{xy} = \frac{\sum_{i=1}^{n}[(x_i - \mu_x)(y_i - \mu_y)]}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2 * \sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

Here, r_{xy} is the correlation between an epoch of a user on day x and day y, x_i is the energy consumption in ith interval on day x, y_i is the energy consumption in ith interval on day y, μ_x is the mean energy consumption in that epoch on day x, μ_y is the mean energy consumption in that epoch on day y, i is the number of samples in an epoch (16 in this work).

From the correlation matrix, the heat map is plotted where correlation coefficient values that were closer to 1 (highly correlated) are darker in shade while values that were closer to 0 (uncorrelated) and -1 (reverse correlated) were lighter in shade [15]. The heat map provides a visual representation for understanding the variations in the consumption pattern of a user on different days of the week. These variations play an important role while clustering the load profiles of different types of users for their consumption on different days of the week. Hence EDA done in this work helps in identifying the epoch size for feature extraction and clustering, understanding the distribution of load profile, and the effect of Time of Use and Day of Use on the consumption pattern of a user.

2.2 Principal Component Analysis (PCA)

PCA is a dimensionality reduction, feature extraction technique where input variables are combined in a way that the “least important” variables can be dropped while retaining the important ones. It is easier and more efficient to explore and visualize smaller data while applying various machine learning models [16]. PCA is an orthogonal linear transformation where data is transformed into a new coordinate system that computes the vector having the largest variance on the first coordinate, the second largest variance on the second coordinate, and so on [17]. Let there be a data set X arranged in the form of a matrix with n rows containing samples and p columns containing different variables for each sample. After column-wise standardization of the matrix in a way that mean of X is 1, it is transformed to an f-dimensional feature space, W, such that $f << p$. The new feature set Y is defined as:

$$y_i = W^T x_i$$

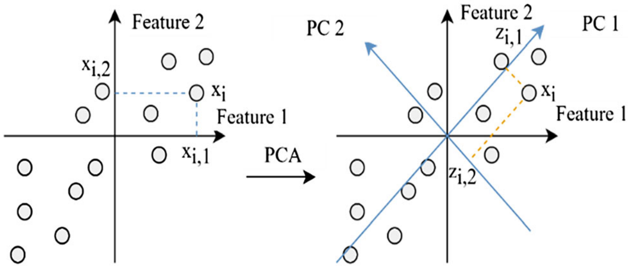


Figure 2: Principal components representing variance in two dimensions

where $i = 1, 2, \dots, n$.

The columns in matrix W are the eigenvalues e_i obtained from the eigenvalue decomposition of:

$$\lambda_i e_i = Q e_i$$

where λ_i is the eigenvalue associated with the eigenvector e_i and Q is the covariance matrix of data set X.

Figure 2 gives the directional representation of principal components where the axes are rotated in a way that they are orthogonal to each other and the variance of the data point projected on the principal component axis is maximized.

Now, to identify the number of principal components carrying maximum information, "Explained Variance" is used. It is a measure of the variation attributed to each principal component in a dataset [16]. It ranks different components in a way that the most important principal components containing maximum variance can be identified. These components are then taken as input to the ML model. It is the ratio of the respective eigenvalue with the sum of all eigenvalues of all eigenvectors.

2.3 Clustering

Clustering is an unsupervised machine-learning technique where similar entities in a dataset are assigned to one cluster. It is an iterative process of multi-object optimization. Its goal is to find high-quality clusters so that the inter-cluster similarity is low and intra-cluster similarity is high [18,19]. In this work, clustering is done to group the users showing similar characteristics in one group. Since the use of electricity is dependent upon "time of use" and "day of use", the clustering is done at an epoch level for each user. In this work partitional clustering algorithm *i.e.* k-means clustering is used to cluster similar load profiles. It aims to partition n data points into k clusters such that each data point belongs to the cluster with the nearest mean.

K means clustering aims to cluster n data points (x_1, x_2, $x_3, \dots, x_n$) into k ($\leq n$) groups so that within the cluster sum of square (WCSS) is minimized. Following are the steps to perform k means clustering:

1. Firstly, a random selection of k cluster centers from the data set is done.
2. Then Euclidean distance is found for each cluster center from each observation and each data point is associated with the cluster giving minimum distance. This form $S_1, S_2, \dots, S_k$ clusters.
3. Then, the cluster centers are updated with the mean value of observation in the current cluster set.
4. Steps 2 and 3 are iterated until cluster sets are fixed and no change is observed.

To converge to local minima the following expression is used:

$$\arg\min \sum_{i=1}^{k} \sum_{x \epsilon S_i} ||x - \mu_i||^2$$

Here, μ_i is the mean of points in cluster Si and x is the corresponding data point.

To find the number of clusters two approaches are used and the intersection of results from both approaches is taken for identifying the clusters. To find the optimal value of "k" the first method is the "Elbow Method". In this work, WCSS is found in clusters between 0 and 15. WCSS is the sum of the squared distance between each point and centroid in a cluster. As the number of clusters increases, the WCSS value decreases and reaches a point where the graph rapidly changes creating an "elbow" shape. The "k" value corresponding to this is considered optimal for the number of clusters [20]. The second approach used to validate the number of clusters is by calculating Silhouette Coefficient. It finds the intra-cluster similarity and inter-cluster dissimilarity by calculating the average distance. Its value ranges between −1 and +1 where −1 indicates weak clustering and +1 indicates clusters are well apart from each other.

3. WORK DONE

This work aims to create cluster groups such that different residential customers and small and medium enterprise (SME) users showing similar consumption characteristics are grouped in one cluster. This helps in demand response management in a way that utilities can analyze the electricity consumption of users falling in one cluster and target them for flattening of peak load by incorporating techniques like dynamic pricing. For understanding the load profile consumption patterns and analyzing and extracting their key features "Irish Smart Meter Database" is used [21].

The Irish database used in this work is robust data containing data samples of around 1.5 years taken from close to 7000 smart meter users. These users include both residential and SME consumers showing different load profile characteristics. Out of these, 1415 users are randomly selected after pre-processing and cleaning of data (removal of outliers and users with missing values). For these 1415 users, two months of data are taken for analysis. The data is sampled at a half-hourly frequency; hence, we get 48 data points for each user in a day. There are around 4.2 million samples taken for analysis. Clustering of such a large dataset increases the computation cost hence feature extraction is done to reduce the data in a way that key features are extracted from the dataset in each epoch and the remaining are removed. Before extracting the features, it is essential to understand the distribution of data. Hence, exploratory data analysis is done for understanding the distribution of data and identifying the optimal epoch size. Figure 3 shows the load profile characteristics of 10 randomly selected residential and SME users over weekdays. The dotted rectangular box on this figure shows data samples of a day. It can be observed that in the first 8 h of a day, the consumption of electricity is lesser and consistent for all users whereas, in the next 8 h of a day, this consumption increases with some peaks. To further analyze such characteristics of load profiles, EDA is done as explained in the next section.

3.1 EDA

In Figure 3 domestic users show a similar load profile on weekdays while SME users (ID 5985) peak on 3 days of the week with negligible power consumption on the remaining days. It can also be seen that in a day (containing 48-time samples) the consumption pattern varies and thus it is essential to understand the patterns in this variation.

In this work, based on the consumption pattern, data points in a day are divided into 3 epochs and features are calculated in each of these epochs. Figure 4 shows the box plot where data distribution variation is analyzed for each user in each epoch. The figure shows the box plot for epochs 1 and 3. This plot explains the Time of Use characteristics of different users. For example, user ID 5985 peaks in the third epoch and shows skewed distribution while in the first two epochs, it shows negligible consumption of electricity. The distribution is skewed in some cases while normal in others for the same user IDs in a day.

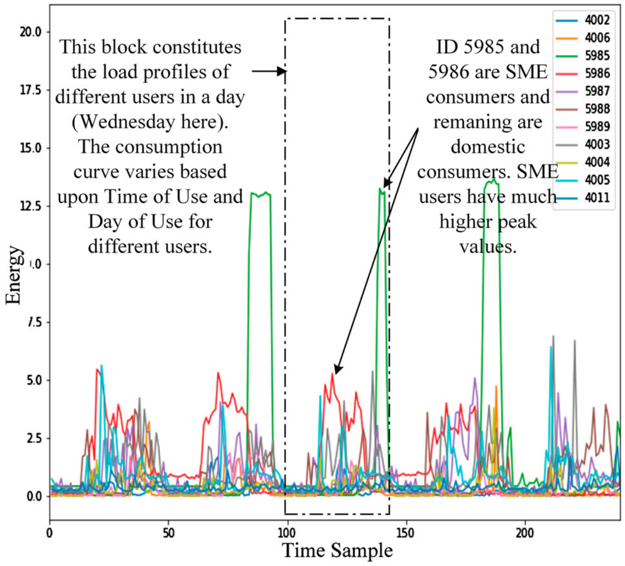


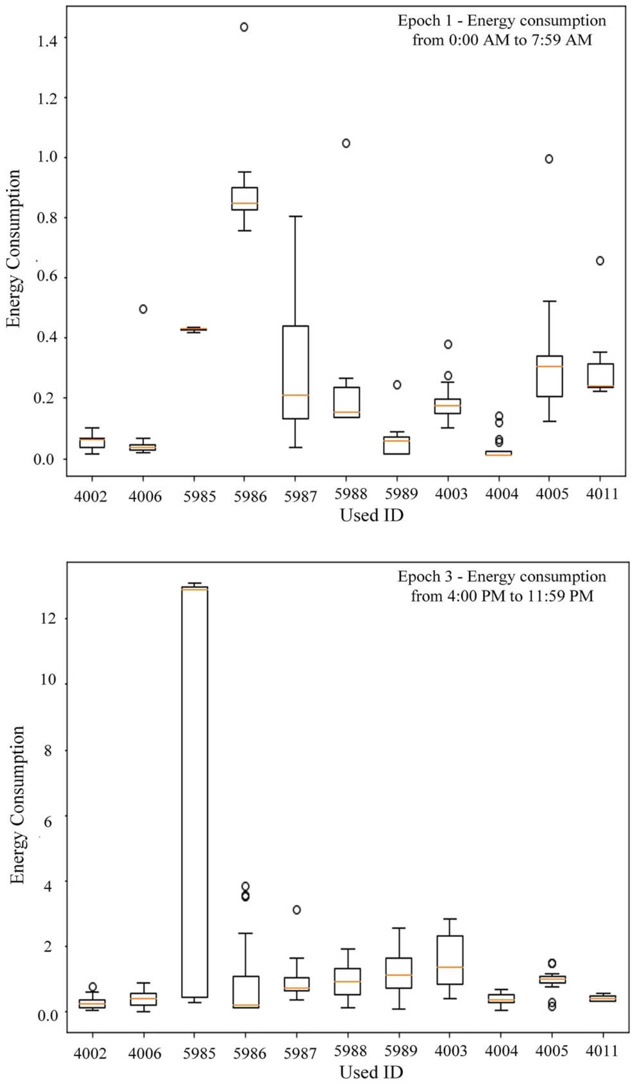
Figure 3: Load profile having total energy consumption in kWh in 30 min for randomly selected 10 SME and domestic consumers for a period of 5 weekdays (December 2009)

To understand the Day of Use characteristics, inter-day correlation is found between the epochs of a user. Figure 5 shows the heat map of user ID 5985 where inter-day correlation is observed to be higher in epoch 2 while lesser in epoch 1 and epoch 3 on day 1 while on days 2, 3, and 4 the correlation is higher in epoch 3 and its value is closer to 1. Thus, it is essential to consider different days of the week while clustering the load profiles so that all variations and conditions are considered.

The variations in load profiles analyzed through EDA play an important role in the clustering of different users. For clustering, it is essential to reduce the data set and extract maximum information from the extracted features. For extracting features taking the correct epoch size is essential. With the help of EDA, it is explained that the epoch size of 16 samples per epoch should be considered for feature extraction and clustering, as it incorporates the variation in the load profile of users while maintaining its correlation with different days of a week. For Feature Extraction PCA is used in this work as explained in the next section.

3.2 Feature Extraction

The dataset used in this work contains around 4.2 million electricity consumption data points. For extracting features and reducing the overall size of the data it is



SME Users - 5985, 5986
Domestic Users - 5987, 5988, 5989, 4003, 4004, 4005, 4011
Unclassified Users - 4002, 4006

Figure 4: Load profile having total energy consumption in kWh in an interval of 30 min for randomly selected 10 SME and domestic consumers for a period of 5 weekdays (December 2009)

arranged in a way that each row contains consumption data in an epoch (arranging the data as 0.263 million $*$ 16). Each sample in an epoch is considered a variable and it is standardized by subtracting the mean and dividing by the standard deviation. This dataset was then used for PCA.

This work aims to extract the principal components in a way that the important information is not lost in dimensionality reduction. In this work principal components are found in each epoch as per the following steps –

- Extract principal components in each epoch to find the number of principal components capturing maximum variance.
- Identify the cumulative variance captured by the components in each epoch and check the number of components required for capturing 95% of data variance (threshold set in this work). Figure 6 shows the cumulative variance of energy consumption of a user in an epoch plotted against the number of principal components. The value of explained variance indicates the total variance captured by including a given number of principal components. It can be seen from the figure that 98% variance is captured by 6 principal components. Iterations are done for up to 15 principal components and it is found that when PC $= 6$ then the average cumulative variance is around 95% for all users.
- Calculate PCA and take the reduced data set with 6 principal components in each epoch as input to the clustering model. With the help of PCA, data set is reduced from 0.263 million $*16$ points to 0.263 million $*$ 6 data points.

PCA leads to a reduction of size of data by around 64%. This reduces the overall computational cost of the ML model while clustering. The time taken to cluster the load profiles using transformed data is halved. Apart from this, it is also observed that feature extraction through PCA leads to a higher average silhouette coefficient as compared to the statistical features calculated earlier [11,22]. Silhouette coefficient determines the inter-cluster dissimilarity and intra-cluster similarity. Hence, the technique used in this work reduces the dimension of the dataset in such a way that the data essential for effective clustering of user load profiles is retained.

4. RESULTS AND DISCUSSION

By using PCA based feature extraction technique, the feature set of 4.2 million samples of two months' electricity consumption data of around 1405 users is reduced to 1.5 million samples. This is done to reduce the computational cost and overall time in the clustering process. It is also done so that well-separated clusters are formed. In this work centroid-based clustering technique *i.e.* "k means" is used for grouping the load profiles showing similar characteristics. It is found that clustering after feature extraction gives well-separated clusters and reduces the computation time by half when compared with the clustering of the raw dataset. To find the optimal value of k for clustering the load profiles, iterations are done from $k = 1$ to $k = 15$ and the within-cluster sum of squares (WCSS) is found for both SME and residential users. Figure 7 shows the WCSS values with respect to the increase in the number of clusters of residential users. From the plot, it is observed that at $k = 5$ and $k = 6$ there

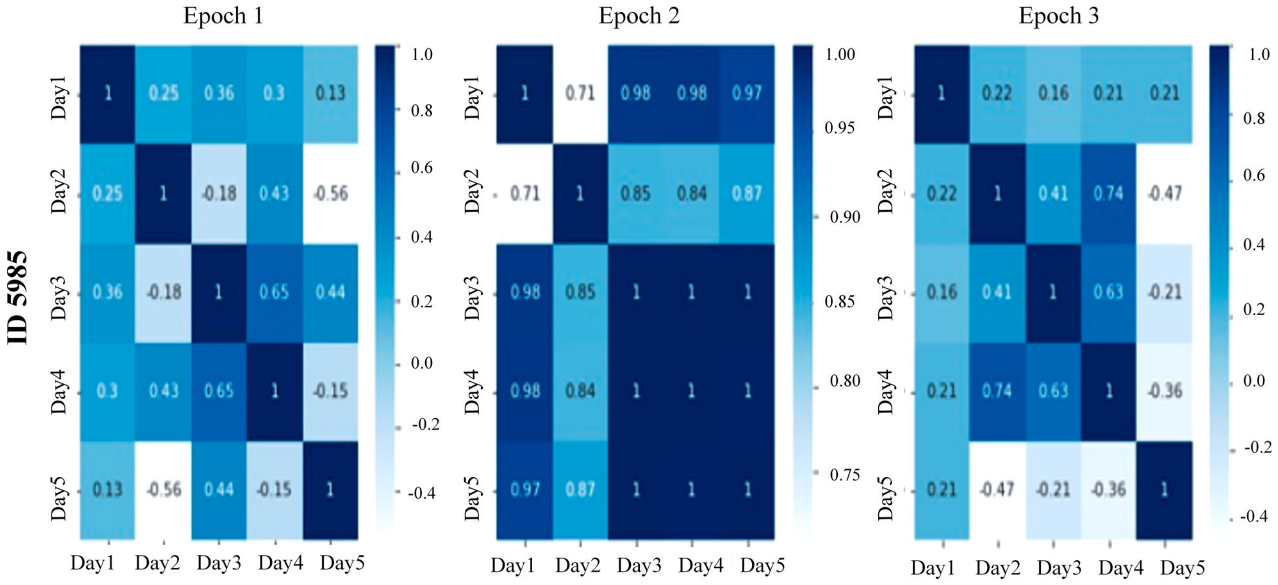


Figure 5: Heat map showing correlation between different days of a week in each epoch of a user

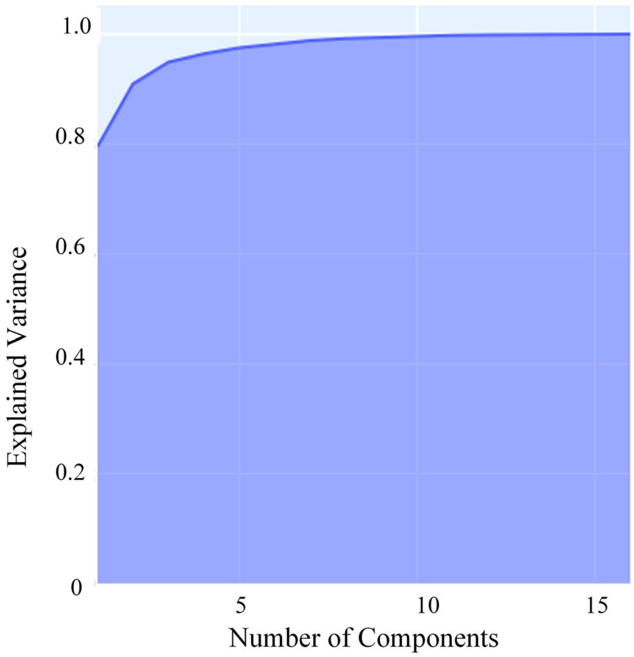


Figure 6: Per unit cumulative variance of electricity consumption of a user on increasing the number of principal components in an epoch

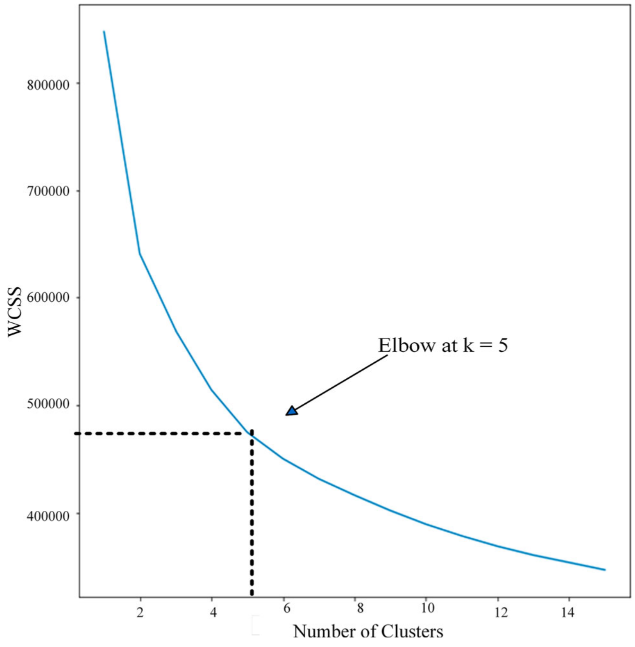


Figure 7: WCSS values with respect to number of clusters showing that $k = 5$ and $k = 6$ will give optimal number of clusters

is an elbow that represents a sudden change in the WCSS values. Hence these are taken while clustering the data. Similarly for SME users also WCSS values are plotted and $k = 3$ is found to be the optimal value for clustering.

For identifying whether the value of “k” should be 5 or 6, the average silhouette coefficient is evaluated for different values of k and it is found that for $k = 5$, the value is highest which indicates that the inter-cluster distance is maximum while intra-cluster distance is minimum for this value of k. Table 1 shows average silhouette coefficient values for each value of “k” for SME and residential users of electricity. For SME users, the silhouette value is 0.511 (highest at $k = 3$) whereas, for residential users, it is 0.480 (highest at $k = 5$). It is observed from these values that the consumption pattern of SME users is similar among themselves with consistent peaks in epoch 2 or epoch 3 whereas for residential users the consumption pattern is dissimilar for different types of users and hence their characteristics are further analyzed as shown

Table 1: Clustering results for optimal value of *k*

	Average silhouette value	
Number of clusters	SME	Residential
3	**0.511**	0.316
4	0.460	0.447
5	0.400	**0.480**
6	0.396	0.340

Table 2: Post clustering analysis of residential users

Cluster Epoch	1	2	3	4	5
Average Energy Consumption of all users in a cluster (kWh)					
1	10.19	3.77	8.88	21.33	6.11
2	16.69	7.01	15.78	30.27	7.98
3	21.75	11.69	20.10	33.47	13.56
Total users in a cluster in an epoch					
1	742	1105	942	316	289
2	1057	1107	1089	845	838
3	1061	1102	1085	794	964

in Table 2. It can be observed from Table 2 that the average energy consumption in cluster 4 is the highest and the number of users falling in this cluster in epoch 2 and epoch 3 are also high. Hence these users are contributing to peaks and can be targeted for demand response management. Similar observations are identified for SME users. This analysis can further be segregated at the day level for identifying the days of the week on which maximum peaks are observed. Although the average consumption of electricity for some clusters are very close to each other (*e.g.* cluster 3 epoch 3 and cluster 1 epoch 3), but due to differences in their patterns in an epoch they are assigned to different clusters. And that is why feature extraction is necessary to incorporate the pattern s variations while clustering the load profiles.

Figure 8 represents the clustering results for $k = 5$ and it is observed that 5 well-separated clusters are formed with this clustering technique. The plot in Figure 8 shows the first two principal components as they carry the maximum variance present in the data. The 5 clusters formed for this dataset for residential users are well separated from each other with centroids lying within the cluster to minimize the intra-cluster distance of all the data points lying in the respective cluster. Each epoch is assigned to a cluster based on its load profile. The value of the average silhouette coefficient ranges from -1 to $+1$. The higher the value better is the clustering result. In this work, the coefficient value is close to 0.5 and hence it is inferred that the clusters are well separated from each other, hence, they can be used for identifying similar load profiles and implementing dynamic pricing based on the consumption pattern.

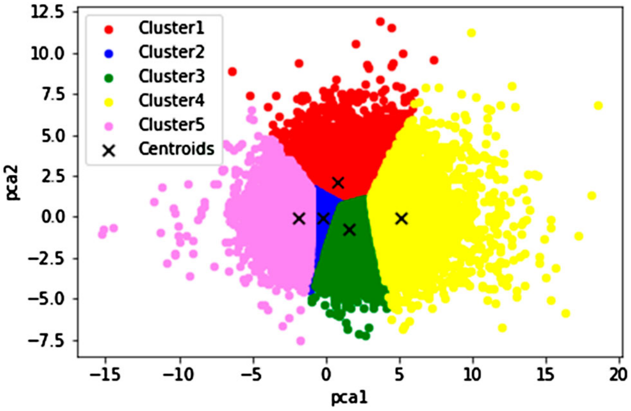


Figure 8: *k*-Means clustering result for $k = 5$

5. CONCLUSION

Analysis of high-frequency smart meter data is required for finding the key insights necessary for the reliable operation of smart grids especially load prediction and demand response management. In this work, a technique is suggested for clustering similar load profiles so that the results can be used for targeting the users eligible for demand response management. "Irish Smart Meter Database" is used for analyzing and implementing the Machine Learning feature extraction and clustering techniques. As it is essential to understand the distribution of data before implementing any ML model exploratory data analysis (EDA) is carried out. It is justified in this work through EDA that taking an epoch of 16 samples at a time for feature extraction rather than the data of the entire day is a comparatively more optimal approach. It is also shown that the Time of Use and Day of Use characteristics should be considered while clustering the load profiles. For feature extraction, PCA is used in this work taking 6 principal components as features (identified by cumulative variance). The dataset is reduced by around 64% using the PCA feature extraction technique. It is observed that the results of k-means clustering are better and faster through this method as compared to clustering of raw dataset. The optimal number of clusters ($k = 5$ for residential users and $k = 3$ for SME users) are found by using WCSS based "Elbow method" and by finding the average silhouette coefficient.

In the future, the results of this model will be used for training the ML models for predicting load demand in each cluster group. This will help utilities in identifying the load demand and hence taking the required actions accordingly. This can also play an important role in the flattening of load curves and the overall conservation of electricity.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

REFERENCES

1. Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Trans. Smart Grid*, Vol. 10, no. 3, pp. 3125–48, 2019. DOI: 10.1109/tsg.2018.2818167.

2. M. Azaza, and F. Wallin, "Smart meter data clustering using consumption indicators: Responsibility factor and consumption variability," *Energy Procedia*, Vol. 142, pp. 2236–42, 2017. DOI: 10.1016/j.egypro.2017.12.624.

3. F. Wang, *et al.*, "Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns," *Energy Convers. Manage.*, Vol. 171, pp. 839–54, 2018.

4. J. Yang, *et al.*, "A model of customizing electricity retail prices based on load profile clustering analysis," *IEEE Trans. Smart Grid*, Vol. 10, no. 3, pp. 3374–86, 2019.

5. A. Rajabi, *et al.*, "A comparative study of clustering techniques for electrical load pattern segmentation," *Renew. Sustain. Energy Rev.*, Vol. 120, pp. 109628, 2020.

6. X. Zhang, *et al.*, "Electricity consumption pattern analysis beyond traditional clustering methods: A novel self-adapting semi-supervised clustering method and application case study," *Appl. Energy*, Vol. 308, pp. 118335, 2022.

7. T. Bomfim, "Evolution of machine learning in smart grids," in *2020 IEEE 8th International Conference on Smart Energy Grid Engineering (SEGE)*, Oshawa, ON, Canada, 2022.

8. Z. A. Khan, D. Jayaweera, and M. S. Alvarez-Alvarado, "A novel approach for load profiling in smart power grids using smart meter data," *Electr. Power Syst. Res.*, Vol. 165, pp. 191–8, 2018. DOI: 10.1016/j.epsr.2018.09.013.

9. E. Lee, J. Kim, and D. Jang, "Load profile segmentation for effective residential demand response program: Method and evidence from Korean pilot study," *Energies*, Vol. 13, no. 6, pp. 1348, 2020.

10. A. Tureczek, P. Nielsen, and H. Madsen, "Electricity consumption clustering using smart meter data," *Energies*, Vol. 11, no. 4, pp. 859, 2018.

11. A. M. Alonso, F. J. Nogales, and C. Ruiz, "Hierarchical clustering for smart meter electricity loads based on quantile autocovariances," *IEEE Trans. Smart Grid*, Vol. 11, no. 5, pp. 4522–30, 2020.

12. P. Nystrup, *et al.*, "Clustering commercial and industrial load patterns for long-term energy planning," *Smart Energy*, Vol. 2, pp. 100010, 2021.

13. Z. Niu, J. Wu, X. Liu, L. Huang, and P. Nielsen, "Understanding energy demand behaviors through spatio-temporal smart meter data analysis," *Energy*, Vol. 226, pp. 120493, 2021. DOI: 10.1016/j.energy.2021.120493.

14. A. Zakariazadeh, "Smart meter data classification using optimized random forest algorithm," *ISA Trans.*, Vol. 126, pp. 361–9, 2021. DOI: 10.1016/j.isatra.2021.07.051.

15. S. Yan, *et al.*, "Time–frequency feature combination based household characteristic identification approach using smart meter data," *IEEE Trans. Ind. Appl.*, Vol. 56, no. 3, pp. 2251–62, 2020. DOI: 10.1109/tia.2020.2981916.

16. P. Bruce, A. Bruce, and P. Gedeck. *Practical Statistics for Data Scientists*. Sebastopol: O'Reilly Media, Incorporated, 2020.

17. G. Reddy, *et al.*, "Analysis of dimensionality reduction techniques on big data," *IEEE. Access.*, Vol. 8, pp. 54776–88, 2020. DOI: 10.1109/access.2020.2980942.

18. L. Kuncheva, and W. Faithfull, "PCA feature extraction for change detection in multidimensional unlabeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, Vol. 25, no. 1, pp. 69–80, 2014. DOI: 10.1109/tnnls.2013.2248094.

19. A. Al-Wakeel, and J. Wu, "K-means based cluster analysis of residential smart meter measurements," *Energy Procedia*, Vol. 88, pp. 754–60, 2016. DOI: 10.1016/j.egypro.2016.06.066.

20. C. Yuan, and H. Yang, "Research on K-value selection method of K-means clustering algorithm," *J*, Vol. 2, no. 2, pp. 226–35, 2019. DOI: 10.3390/j2020016.

21. Irish Social Science Data Archive, "Commission for energy regulation (CER) smart metering project." 2012. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

22. G. Shamim, and M. Rihan, "Novel technique for feature computation and clustering of smart meter data," in *2019 International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Aligarh, 2019.

AUTHORS

Gulezar Shamim received her Btech degree and Mtech degree in electrical engineering from Aligarh Muslim University in 2013 and 2015 respectively. She is currently pursuing her PhD in the department of electrical engineering, AMU, India. Her research interests are with the field of big data analytics in smart grids.

Corresponding author. Email: gulezar2009@gmail.com

Mohd Rihan received his Btech, Mtech and PhD in electrical engineering from Aligarh Muslim University in 2001, 2005 and 2013 respectively. He has about 20 years of teaching experience in AMU and is the founder coordinator of Centre for Grid Integrated Green and Renewable Energy in the campus and a member secretary of Central Project Implementation and Monitoring Committee. He is an elected fellow of IET (UK), IE (India) and IETE and a senior member of IEEE. His research interests are in the field of smart grids and solar energy.

Email: m.rihan.ee@amu.ac.in