

Clustering Prosumers' Data to Identify their Demand Response potential in Smart Grids

submitted by

Ankita Dutta

Class Roll No: 002210503030

Exam Roll: MCA244029

Master of Computer Application

under the supervision of

Prof. (Dr.) Ujjwal Maulik

in the

Department of Computer Science and Engineering

Faculty of Engineering and Technology

Jadavpur University

2024

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF RECOMMENDATION

To Whom It May Concern

I hereby recommend that the project report entitled “Clustering Prosumers' data to identify their Demand Response potential in Smart Grids” has been carried out by Ankita Dutta (Reg. No.: 163622 of 2022-23, Roll No.: 002210503030, Exam Roll: MCA244029), under my guidance and supervision and be accepted in partial fulfilment of the requirement for the degree of Master of COMPUTER APPLICATION in DEPARTMENT of COMPUTER SCIENCE and ENGINEERING, JADAVPUR UNIVERSITY during the academic year 2023-2024.

Prof. (Dr.) Ujjwal Maulik
Project Supervisor
Dept. of Computer Science and
Engineering
Jadavpur University

Prof. (Dr.) Nandini Mukhopadhyay
Head of the Department
Dept. of Computer Science & Engineering
Jadavpur University

Prof. Dipak Laha
Dean, Faculty Council
of Engineering & Technology
Jadavpur University, Kolkata-032

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF APPROVAL

This is to certify that the project report entitled “**Clustering Prosumers' data to identify their Demand Response potential in Smart Grids**” is a bonafide record of work carried out by ANKITA DUTTA, in fulfillment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the project report only for the purpose for which it has been submitted.

Signature of Examiner 1

Date :

Signature of Examiner 2

Date :

Faculty of Engineering and Technology
Jadavpur University

**DECLARATION OF COMPLIANCE WITH
ACADEMIC ETHICS**

I hereby declare that this project report entitled "Clustering Prosumers' data to identify their Demand Response potential in Smart Grids" is an outcome of my genuine and independent work, as part of my Master of Computer Application program.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, all materials and results that are not original to this work have been fully cited and referenced.

Ankita Dutta

Registration No.: 163622 of 2022-23

Examination Roll No.: MCA244029

Thesis Title :- Clustering Prosumers' data to identify their Demand Response potential in Smart Grids

Signature with Date

ACKNOWLEDGEMENT

With my deepest gratitude, I would like to thank Prof. (Dr.) Ujjwal Maulik of the Department of Computer Science & Engineering, my supervisor, for his unwavering support throughout the duration of this project. His encouragement provided the necessary momentum and inputs to continue my work, and without his guidance, the project would not have taken its current shape. His valuable suggestions and numerous discussions have always inspired new ways of thinking. I feel deeply honored to have had the opportunity to work under his supervision.

I would also like to express my sincere thanks and greatest gratitude to Ms. Kakuli Mishra, who, despite her busy schedule, selflessly guided me throughout the entire project and was always there for me whenever I needed support.

My heartfelt thanks go to all the faculty members of the Department of Computer Science & Engineering at Jadavpur University for their continuous support.

Lastly, but certainly not least, I would like to thank my batchmates for standing by my side when I needed them the most.

Ankita Dutta

Class Roll No.: 002210503030

Examination Roll No.: MCA244029

Department of Computer Science & Engineering

Jadavpur University

Abstract

In modern energy systems, the rise of prosumers—individuals or households that both consume and produce electricity—presents opportunities and challenges for efficient energy management. This study explores using advanced machine learning techniques to cluster prosumers and analyze their consumption and production patterns to enhance demand response (DR) programs. We utilized two datasets: the Ausgrid dataset, containing data from 300 solar-equipped households in Australia, and the Danish dataset, which provides detailed electricity consumption data.

For the Ausgrid dataset, we preprocessed the data, extracted relevant features, and applied clustering algorithms like K-Means, Fuzzy C-Means, K-Medoids, Hierarchical Clustering, and Spectral Clustering. The results, evaluated using metrics such as the Silhouette Score and Davies-Bouldin Index, revealed distinct clusters of prosumers with similar energy patterns, offering insights for targeted DR programs.

In the Danish dataset, we used subsequence clustering to identify recurring patterns within the time series data. This involved dividing the time series into overlapping subsequences and clustering them. The analysis showed minimal seasonal influence on the clusters, suggesting other factors significantly shape energy behaviors.

These insights have practical implications for energy utilities and policymakers. Understanding prosumer patterns can help develop effective DR programs, enhancing grid stability and energy savings. The minimal seasonal impact on energy patterns suggests that DR strategies could be universally applicable year-round.

This study demonstrates the applicability of machine learning techniques to clustering analysis in energy management, providing a blueprint for optimizing energy systems in a prosumer-dominated landscape.

Contents

Abstract	i
Table of Contents	ii
List of Figures	v
List of Tables	vii
1 Introduction	1
1.1 Machine Learning	2
1.1.1 Supervised Learning	2
1.1.2 Unsupervised Learning	3
1.1.2.1 Clustering Algorithms	3
1.1.2.2 Dimensionality Reduction	7
1.2 Time Series Clustering	11
1.2.1 Prosumer Clustering	12
1.2.1.1 Importance of Prosumer and Daily Consump-	
tion Pattern Clustering	12
1.2.1.2 Demand Response (DR) Programs	12
1.2.1.3 Clustering of Customers	13
1.2.1.4 Clustering of Daily Consumption Patterns .	13
1.2.2 Evaluation Metrics	13
1.2.2.1 Silhouette Score	14
1.2.2.2 Calinski Harabasz Index	14
1.2.2.3 Davies Bouldin Index	15
1.2.2.4 Intercluster Distance	15
1.2.2.5 Intraccluster Distance	16
1.2.2.6 Adjusted Rand Index	16
1.2.2.7 Adjusted Mutual Information	16
1.2.2.8 Entropy	17
1.3 Clustering Methodology	18
1.3.1 Data Cleaning	18
1.3.2 Feature Extraction	18
1.3.3 Normalization	19
1.3.4 Clustering Algorithm Selection	19
1.3.4.1 Finding optimal number of clusters	19
1.3.4.2 Identifying best performing algorithm	20
1.3.5 Visualization	20

2	Literature Survey	21
2.1	Prosumer Clustering in Smart Grids	21
2.2	Management of Prosumers in Smart Grids	23
2.3	Time Series Clustering Techniques	24
2.4	Applications of Clustering in Smart Grids	24
2.5	Clustering Residential Electricity Customers	25
2.6	Time Series Clustering Methods	26
2.7	Machine Learning-Based Clustering	27
2.8	Probabilistic Load Flow Approaches	27
2.9	Distributed Energy Resources Coordination	28
2.10	Energy Sharing and Optimization	28
2.11	Solar Prosumer Identification	29
3	Prosumer Clustering with Ausgrid Dataset	31
3.1	Dataset Overview	31
3.2	Cluster Analysis	32
3.3	Interpretation of Gross Generation Capacity	35
	3.3.0.1 Clustering Results	36
	3.3.0.2 Visualization of Clusters	36
	3.3.0.3 Discussion on Visualizations	36
4	Prosumer Clustering with Synthetic Dataset	47
4.1	Dataset Overview	47
4.2	Clustering Analysis	48
5	Future Work and Conclusion	59
	Bibliography	63

List of Figures

1.1	Proposed Methodology for Prosumer Clustering	18
3.1	Silhouette Score Comparison	33
3.2	Davies-Bouldin Index Comparison	34
3.3	Calinski-Harabasz Index Comparison	35
3.4	Time-Series Patterns of the Three Clusters for Gross Generation (GG)	38
3.5	Time-Series Patterns of the Three Clusters for General Consumption (GC)	39
3.6	Time-Series Patterns of the Three Clusters for Controlled Load (CL)	40
3.7	PCA Scatter Plot of KMeans Clusters	41
3.8	Surplus/Deficit Energy by Cluster	41
3.9	Sum of Squared Distances (SSD) vs Number of Clusters (k) for GG Data	42
3.10	PCA Plot of GG Clusters	42
3.11	GG Patterns of the Three Clusters	43
3.12	Average GG Energy by Cluster	44
3.13	Peak GG for Each Member of Each Cluster	44
3.14	GG Line Graphs of All Customers with Medians Distinguishing Each Cluster	45
4.1	Silhouette Score vs Number of Clusters for Different Algorithms	49
4.2	GG Time Series Data by Cluster	51
4.3	GC Time Series Data by Cluster	52
4.4	Davies-Bouldin Score vs Number of Clusters for Different Algorithms	53
4.5	Calinski-Harabasz Score vs Number of Clusters for Different Algorithms	54
4.6	PCA Plot of KMeans Clusters	55
4.7	Daywise Distribution of KMeans Clusters across Seasons	56
4.8	Count of Each Season in Each KMeans Cluster	57

List of Tables

3.1	Description of the columns in the Ausgrid dataset.	32
3.2	Evaluation Scores and Distances for Clustering Algorithms ($k = 3$)	33
3.3	Evaluation Scores and Distances for Clustering Algorithms ($k = 3$)	36
4.1	Evaluation Scores and Distances for Clustering Algorithms ($k = 3$)	48

CHAPTER 1

Introduction

Machine learning (ML) has become a cornerstone of modern data analysis, transforming how we approach various problems across numerous industries. This paper focuses on utilizing ML techniques for clustering prosumers, specifically in the context of electricity consumption. Prosumers, individuals or households that both consume and produce electricity, represent a unique subset of energy users whose behavior can significantly impact demand response (DR) programs. By clustering these prosumers based on their consumption and generation patterns, utilities can design more efficient and targeted DR strategies, ultimately contributing to better energy management and sustainability.

This research examines two primary datasets: the Ausgrid dataset, which includes data from 300 solar-equipped households in Australia, and the Danish dataset, which encompasses broader and more detailed electricity consumption data. The primary goal is to apply ML techniques to cluster these prosumers effectively, providing insights that can enhance DR programs. For the Ausgrid dataset, we focus on clustering customers based on their overall consumption and generation patterns. For the Danish dataset, the focus shifts to subsequence clustering, identifying recurring patterns within the time series data.

By understanding the consumption and generation patterns of prosumers, utilities can design more effective programs that encourage energy savings and improve grid stability. The methodologies employed in this study include K-Means, hierarchical clustering, Fuzzy C-Means (FCM), K-Medoids, and spectral clustering, along with dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). These methods are evaluated using metrics such as the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score.

In the following sections, we will first provide a foundational understanding of machine learning, exploring its different types, particularly supervised and unsu-

pervised learning. We will then delve into time series clustering, a crucial technique for analyzing data that changes over time, such as electricity consumption patterns. Finally, we will discuss prosumer clustering, its significance in the context of DR programs, and the specific methodologies employed in this research.

1.1 Machine Learning

Machine learning (ML) is a subset of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to perform specific tasks without explicit instructions. Instead, these models learn from data, identifying patterns and making decisions based on their analysis. The primary objective of ML is to develop models that can generalize from the training data to new, unseen data, providing accurate predictions or insights. ML can be broadly classified into three categories: supervised learning, unsupervised learning, and reinforcement learning. This paper focuses on supervised and unsupervised learning, as they are directly applicable to the clustering techniques employed in our analysis.

1.1.1 Supervised Learning

Supervised learning is a type of machine learning where the model is trained on a labeled dataset. This means that each training example is paired with an output label. The goal of the supervised learning algorithm is to learn a mapping from the input data to the output labels, such that it can accurately predict the label for new, unseen data. Supervised learning can be further categorized into two main types: classification and regression.

- **Classification:** Classification tasks involve predicting a discrete label or category for each input data point. For example, an email spam filter is a classification model that classifies emails as either "spam" or "not spam." In the context of electricity consumption, a classification model might predict whether a given consumption pattern indicates normal usage or an anomaly.
- **Regression:** Regression tasks involve predicting a continuous value for each input data point. An example of a regression task is predicting house prices based on features such as size, location, and number of bedrooms. In energy consumption analysis, regression models can be used to predict future electricity usage based on historical consumption data and other relevant factors.

Supervised learning models include a variety of algorithms, each with its strengths and weaknesses. Common algorithms used in supervised learning include:

1. **Linear Regression:** A simple yet powerful algorithm for regression tasks, where the model predicts the output by fitting a linear relationship between the input variables and the target variable.
2. **Decision Trees:** A versatile algorithm that can be used for both classification and regression tasks. Decision trees split the data into subsets based on feature values, creating a tree-like model of decisions.
3. **Random Forests:** An ensemble method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.
4. **Support Vector Machines (SVM):** A robust algorithm used for classification tasks, SVMs find the optimal hyperplane that separates the data points of different classes with the maximum margin.
5. **Neural Networks:** Inspired by the human brain, neural networks are highly flexible models that can learn complex patterns in data. They are particularly useful for tasks with large and complex datasets.

In the context of our research, while supervised learning is not the primary focus, understanding its principles is essential as it forms the foundation for many advanced ML techniques and provides a contrast to the unsupervised learning methods used for clustering. By mastering these algorithms and their applications, we can better appreciate the strengths and limitations of the various clustering techniques employed in this study.

1.1.2 Unsupervised Learning

Unsupervised learning involves training a model on data without labeled responses. The goal is to infer the natural structure present within a set of data points. Unlike supervised learning, there are no output labels to guide the learning process. Unsupervised learning is particularly useful for exploratory data analysis, pattern recognition, and clustering.

1.1.2.1 Clustering Algorithms

Clustering is a type of unsupervised learning where the objective is to group a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. This technique is widely used

in various fields, including market research, social network analysis, image segmentation, and, as in our case, electricity consumption analysis. Below are detailed explanations of the clustering algorithms used in this research: K-Means, Fuzzy C-Means (FCM), K-Medoids, Hierarchical Clustering, and Spectral Clustering.

1. **K-Means Clustering:** KMeans[1][2][3] is one of the simplest and most popular clustering algorithms. It aims to partition the data into K clusters by minimizing the variance within each cluster.

The algorithm follows these steps:

- (a) Initialize K centroids randomly.
- (b) Assign each data point to the nearest centroid, forming K clusters.
- (c) Recompute the centroids as the mean of the data points in each cluster.
- (d) Repeat steps 2 and 3 until the centroids no longer change significantly.

The objective function for K-Means is to minimize the sum of squared distances between each data point and its corresponding centroid:

$$J = \sum_{i=1}^K \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1.1)$$

where:

- J : The objective function, which we aim to minimize.
- K : The number of clusters.
- C_i : The set of points (or the cluster) assigned to the i -th centroid.
- x : A data point within the cluster C_i .
- μ_i : The centroid of the i -th cluster.
- $\|x - \mu_i\|^2$: The squared Euclidean distance between a data point x and the centroid μ_i .

2. **Fuzzy C-Means Clustering:** Fuzzy C-Means (FCM)[4][5][6] is a clustering method that allows each data point to belong to multiple clusters with varying degrees of membership. This provides a more nuanced clustering approach compared to K-Means. The steps of FCM are as follows:

- (a) Initialize the membership matrix randomly.
- (b) Compute the cluster centroids using the membership values.

- (c) Update the membership matrix based on the distance of each data point to the centroids.
- (d) Repeat steps 2 and 3 until convergence.

The objective function for FCM is to minimize the weighted sum of squared errors:

$$J_m = \sum_{i=1}^K \sum_{j=1}^N u_{ij}^m \|x_j - \mu_i\|^2 \quad (1.2)$$

where:

- J_m : The objective function, which we aim to minimize.
- K : The number of clusters.
- N : The number of data points.
- u_{ij} : The degree of membership of x_j in cluster i .
- m : The fuzziness parameter, which determines the level of cluster fuzziness.
- x_j : A data point.
- μ_i : The centroid of the i -th cluster.
- $\|x_j - \mu_i\|^2$: The squared Euclidean distance between a data point x_j and the centroid μ_i .

3. **K-Medoids Clustering:** K-Medoids[7][8][9], also known as Partitioning Around Medoids (PAM), is similar to K-Means but instead of using the mean of the cluster, it uses actual data points (medoids) to define the center of each cluster. This makes K-Medoids more robust to noise and outliers. The steps of K-Medoids are as follows:

- (a) Initialize K medoids randomly.
- (b) Assign each data point to the nearest medoid, forming K clusters.
- (c) For each cluster, select the data point that minimizes the sum of distances to all other points in the cluster as the new medoid.
- (d) Repeat steps 2 and 3 until the medoids no longer change significantly.

The objective function for K-Medoids is to minimize the sum of pairwise dissimilarities between points in the cluster and the medoid:

$$J = \sum_{i=1}^K \sum_{x \in C_i} \|x - m_i\| \quad (1.3)$$

Here's a breakdown of the components of the formula:

- J : The objective function, which we aim to minimize.
- K : The number of clusters.
- C_i : The set of points (or the cluster) assigned to the i -th medoid.
- x : A data point within the cluster C_i .
- m_i : The medoid of the i -th cluster.
- $\|x - m_i\|$: The dissimilarity (or distance) between a data point x and the medoid m_i .

4. **Hierarchical Clustering:** Hierarchical clustering[10][11][12][13] builds a hierarchy of clusters by either progressively merging smaller clusters (agglomerative approach) or progressively splitting larger clusters (divisive approach). The result is often visualized as a dendrogram, showing the relationships between clusters. The steps of agglomerative hierarchical clustering are as follows:

- (a) Start with each data point as a single cluster.
- (b) Find the two closest clusters and merge them into a single cluster.
- (c) Repeat step 2 until all data points are in a single cluster.

The linkage criterion determines the distance between clusters:

- **Single Linkage:** The minimum distance between points in the two clusters.
- **Complete Linkage:** The maximum distance between points in the two clusters.
- **Average Linkage:** The average distance between points in the two clusters.
- **Ward Linkage:** The increase in variance for the cluster being merged.

5. **Spectral Clustering:** Spectral clustering[14][15][16][17] uses the eigenvalues of a similarity matrix to perform dimensionality reduction before clustering in fewer dimensions. This method is particularly effective for identifying clusters with complex shapes and structures. The steps of spectral clustering are as follows:

- (a) Construct a similarity matrix from the data points.
- (b) Compute the Laplacian matrix from the similarity matrix.
- (c) Compute the eigenvalues and eigenvectors of the Laplacian matrix.
- (d) Use the eigenvectors to project the data into a lower-dimensional space.
- (e) Apply K-Means clustering to the projected data.

The similarity matrix W is defined as:

$$W_{ij} = \exp \left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2} \right) \quad (1.4)$$

The Laplacian matrix L is defined as:

$$L = D - W \quad (1.5)$$

where D is the degree matrix.

Here's a breakdown of the components of the formula:

- W_{ij} : The similarity between data points x_i and x_j .
- σ : The bandwidth parameter that controls the neighborhood size.
- L : The Laplacian matrix used for spectral clustering.
- D : The degree matrix, which is a diagonal matrix containing the sum of each row of W .

1.1.2.2 Dimensionality Reduction

Another important area of unsupervised learning is dimensionality reduction, which involves reducing the number of random variables under consideration, simplifying the data without losing significant information. Techniques such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), and Isometric Mapping (Isomap) are commonly used for this purpose.

Dimensionality reduction is crucial for visualizing high-dimensional data and improving the performance of clustering algorithms by eliminating noise and redundant features.

1. **PCA (Principal Component Analysis):** PCA[18][19] is a linear dimensionality reduction technique that seeks to find the directions (principal components) that capture the maximum variance in the data. It projects the

data onto a lower-dimensional subspace while preserving as much variance as possible.

PCA involves the following steps:

- (a) Compute the covariance matrix of the data. The covariance matrix is calculated as:

$$\text{Cov}(X) = \begin{bmatrix} \text{Cov}(a_1, a_1) & \text{Cov}(a_1, a_2) & \cdots & \text{Cov}(a_1, a_n) \\ \text{Cov}(a_2, a_1) & \text{Cov}(a_2, a_2) & \cdots & \text{Cov}(a_2, a_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(a_n, a_1) & \text{Cov}(a_n, a_2) & \cdots & \text{Cov}(a_n, a_n) \end{bmatrix} \quad (1.6)$$

where a_i represents the i -th feature and $\text{Cov}(a_i, a_j)$ is the covariance between features a_i and a_j .

The covariance between two features a_i and a_j is calculated as:

$$\text{Cov}(a_i, a_j) = \frac{1}{N} \sum_{k=1}^N (a_i^{(k)} - \bar{a}_i)(a_j^{(k)} - \bar{a}_j) \quad (1.7)$$

where N is the number of data points, $a_i^{(k)}$ is the i -th feature of the k -th data point, and $\bar{a}_i$ is the mean of the i -th feature.

- (b) Calculate the eigenvectors and eigenvalues of the covariance matrix. These are obtained by solving the following equation:

$$\text{Cov}(X)\mathbf{v} = \lambda\mathbf{v} \quad (1.8)$$

where $\mathbf{v}$ is the eigenvector and λ is the eigenvalue.

- (c) Select the top k eigenvectors corresponding to the largest eigenvalues to form the new subspace.
- (d) Project the original data onto the new subspace formed by the selected eigenvectors.

2. **t-SNE (t-Distributed Stochastic Neighbor Embedding):** t-SNE[20] is a non-linear dimensionality reduction technique particularly useful for visualizing high-dimensional data in lower-dimensional space. It aims to preserve the local structure of the data by modeling pairwise similarities between data points in high-dimensional space and their counterparts in the low-dimensional embedding.

t-SNE involves the following steps:

- (a) Construct a similarity matrix from the data points. The similarity matrix W is defined as:

$$W_{ij} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right) \quad (1.9)$$

where W_{ij} is the similarity between data points x_i and x_j , $\|x_i - x_j\|_2^2$ is the squared Euclidean distance between them, and σ is the bandwidth parameter that controls the neighborhood size.

- (b) Compute the Laplacian matrix from the similarity matrix. The Laplacian matrix L is defined as:

$$L = D - W \quad (1.10)$$

where D is the degree matrix.

- (c) Compute the eigenvalues and eigenvectors of the Laplacian matrix.
 (d) Use the eigenvectors to project the data into a lower-dimensional space.
 (e) Apply K-Means clustering to the projected data.

3. UMAP (Uniform Manifold Approximation and Projection): UMAP[21]

is a non-linear dimensionality reduction technique that emphasizes preserving both local and global structure of the data. It efficiently approximates the manifold structure of high-dimensional data and projects it onto a lower-dimensional space, often producing more interpretable embeddings compared to t-SNE.

UMAP involves the following steps:

- (a) Construct a high-dimensional fuzzy topological representation of the data.
 (b) Optimize low-dimensional embeddings to preserve the topological structure of the fuzzy simplicial set. This is achieved by minimizing the cross-entropy between the high-dimensional and low-dimensional representations of local neighborhoods. The optimization objective is given by:

$$\text{Loss} = \sum_i \sum_{j \in \mathcal{N}_i} w_{ij} \left(D_{ij} - 2 \log \left(\sum_k \exp(-D_{ik}) \right) + 2 \log \left(\sum_k \exp(-D_{ij}) \right) \right)^2 \quad (1.11)$$

where D_{ij} is the distance between data points i and j , $\mathcal{N}_i$ is the set of neighboring data points of data point i , and w_{ij} is the weight assigned to the edge between data points i and j in the fuzzy simplicial set.

- (c) Use stochastic gradient descent (SGD) to minimize the cross-entropy loss function given by Equation (1.11).
- (d) Optionally, refine the embeddings using a smoother Riemannian metric.

4. **Isomap (Isometric Mapping):** Isomap[22][23] is a non-linear dimensionality reduction technique that focuses on preserving the intrinsic geometric structure of the data. It constructs a neighborhood graph based on pairwise distances between data points, computes the geodesic distances on this graph, and then embeds the data into a lower-dimensional space while preserving these distances as much as possible.

Isomap involves the following steps:

- (a) Construct a neighborhood graph based on the k -nearest neighbors of each data point.
- (b) Compute the shortest path distances (geodesic distances) between all pairs of data points on the neighborhood graph. This is typically done using algorithms such as Dijkstra's algorithm.
- (c) Apply classical multidimensional scaling (MDS) to the geodesic distances to obtain the low-dimensional embedding. Classical MDS minimizes the discrepancy between the pairwise distances in the high-dimensional and low-dimensional spaces. The stress function minimized by classical MDS is given by:

$$\text{Stress} = \left(\sum_{i < j} (d_{ij} - \hat{d}_{ij})^2 \right)^{1/2} \quad (1.12)$$

where d_{ij} is the pairwise distance between data points i and j in the high-dimensional space, and $\hat{d}_{ij}$ is the geodesic distance between the corresponding data points in the low-dimensional embedding.

- (d) Optionally, perform dimensionality reduction using classical MDS on the geodesic distance matrix computed in step 2. Classical MDS will embed the data into a lower-dimensional space while preserving the pairwise geodesic distances as closely as possible.

Including a variety of dimensionality reduction techniques provides researchers with a range of options for visualizing and exploring high-dimensional data, allowing for a more comprehensive analysis and interpretation of clustering results.

1.2 Time Series Clustering

Time series clustering is a specialized form of clustering that deals with data points collected or recorded at specific time intervals. Unlike traditional clustering methods that focus on static data, time series clustering aims to identify patterns and similarities in sequences of data points over time. This technique is particularly useful in fields where temporal patterns are critical, such as finance, meteorology, and, as in this study, energy consumption analysis.

1) Characteristics of Time Series Data

Time series data have unique characteristics that distinguish them from other types of data:

1. **Temporal Dependency:** Data points in a time series are dependent on their previous values, meaning that current values can be influenced by past values.
2. **Seasonality:** Many time series exhibit regular, repeating patterns over time, such as daily, weekly, or yearly cycles.
3. **Trend:** Time series data can show long-term upward or downward movements, indicating a trend.
4. **Noise:** Random variations or fluctuations that do not follow a pattern can also be present in time series data.

2) Challenges in Time Series Clustering

Clustering time series data presents several challenges:

1. **Dimensionality:** Time series data are often high-dimensional, with each series containing a large number of data points.
2. **Distance Measures:** Choosing an appropriate distance measure to compare time series is critical. Common measures include Euclidean distance, Dynamic Time Warping (DTW), and correlation-based distances.
3. **Scalability:** Clustering algorithms must be scalable to handle large datasets efficiently.
4. **Interpretability:** The results of time series clustering should be interpretable and actionable, particularly in practical applications such as DR programs.

1.2.1 Prosumer Clustering

Prosumer clustering is a specialized application of clustering techniques focused on individuals or households that both consume and produce electricity. The term "prosumer" is derived from "producer" and "consumer," reflecting the dual role these entities play in the energy ecosystem. Prosumer clustering involves grouping these entities based on their energy consumption and production patterns, providing valuable insights for optimizing energy management and enhancing demand response (DR) programs.

1.2.1.1 Importance of Prosumer and Daily Consumption Pattern Clustering

Clustering prosumers and daily consumption patterns is crucial for several reasons:

- **Enhanced DR Programs:** Understanding the consumption and generation patterns of prosumers allows utilities to design DR programs that are more targeted and effective. Clusters of prosumers with similar behavior can be offered tailored incentives and strategies to reduce peak demand and balance the grid.
- **Improved Energy Management:** Clustering helps identify prosumers with similar daily usage patterns, enabling better forecasting and management of energy resources. This can lead to more efficient grid operations and reduced reliance on non-renewable energy sources.
- **Customer Engagement:** Understanding the behavior of different prosumer clusters helps utilities engage with their customers more effectively, providing personalized recommendations and improving customer satisfaction.

1.2.1.2 Demand Response (DR) Programs

DR programs encourage consumers to adjust their electricity usage in response to supply conditions. These programs include:

- **Time-of-Use (ToU) Pricing:** Offering different electricity prices at different times of the day to encourage consumers to shift their usage to off-peak hours.

- **Critical Peak Pricing (CPP):** Implementing higher prices during periods of peak demand to incentivize consumers to reduce their usage.
- **Direct Load Control (DLC):** Allowing utilities to remotely control certain appliances in consumers' homes to reduce demand during peak periods.
- **Incentive-Based Programs:** Offering financial incentives to consumers who voluntarily reduce their usage during peak periods.

1.2.1.3 Clustering of Customers

Customer clustering involves grouping consumers based on their daily electricity consumption patterns. This helps utilities identify segments of customers with similar behaviors and tailor DR programs accordingly. For example, households with high evening usage might benefit from ToU pricing, while those with flexible usage patterns might be more suitable for CPP programs.

For the Ausgrid dataset, clustering was performed on the overall consumption and generation patterns of the customers. This involved analyzing the half-hourly electricity meter data to identify clusters of customers with similar usage profiles.

1.2.1.4 Clustering of Daily Consumption Patterns

Clustering daily consumption patterns involves grouping daily electricity consumption data based on their similarity in usage patterns. This analysis aims to identify common usage patterns among different days.

For the Danish dataset, clustering focused on grouping the rows representing daily electricity consumption patterns. This involved analyzing daily variations in energy consumption to uncover common usage patterns among different days.

By clustering prosumers based on their behavior and daily consumption patterns, we aim to provide utilities with actionable insights that can enhance the effectiveness of DR programs, improve energy management, and increase customer satisfaction. This research has the potential to contribute to more sustainable and efficient energy systems, benefiting both utilities and consumers.

1.2.2 Evaluation Metrics

Evaluating the performance of clustering algorithms is essential to ensure the quality and interpretability of the resulting clusters. Several metrics are commonly

used to assess clustering performance, each providing a different perspective on the effectiveness of the clustering. This section discusses the key evaluation metrics used in this research: Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index, Intercluster Distance, Intracluster Distance, Adjusted Rand Index (ARI), and Adjusted Mutual Information (AMI).

1.2.2.1 Silhouette Score

The Silhouette Score [24] is a measure of how similar an object is to its own cluster compared to other clusters. It combines ideas of cohesion (how closely related data points in a cluster are) and separation (how well-separated a cluster is from other clusters). For a data point i , the Silhouette Score $s(i)$ is calculated as follows:

1. Compute the mean distance between i and all other points in the same cluster (a).
2. Compute the mean distance between i and all points in the nearest cluster to which i does not belong (b).
3. The Silhouette Score for i is then given by:

$$s(i) = \frac{b - a}{\max(a, b)} \quad (1.13)$$

The Silhouette Score ranges from -1 to 1, where a higher score indicates better-defined clusters.

1.2.2.2 Calinski Harabasz Index

The Calinski-Harabasz Index [25], also known as the Variance Ratio Criterion, evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion. Higher values indicate better-defined clusters. The Calinski-Harabasz Index CH is calculated as follows:

$$CH = \frac{\text{trace}(B_k)}{\text{trace}(W_k)} \cdot \frac{N - k}{k - 1} \quad (1.14)$$

where:

- B_k is the between-cluster dispersion matrix.
- W_k is the within-cluster dispersion matrix.

- N is the total number of data points.
- k is the number of clusters.

1.2.2.3 Davies Bouldin Index

The Davies-Bouldin Index [26] measures the average similarity ratio of each cluster with its most similar cluster. Lower values indicate better clustering. The Davies-Bouldin Index DB is calculated as follows:

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \quad (1.15)$$

where:

- σ_i and σ_j are the average distances of all points in clusters i and j to their respective centroids.
- $d(c_i, c_j)$ is the distance between the centroids of clusters i and j .
- k is the number of clusters.

1.2.2.4 Intercluster Distance

Intercluster Distance [27] measures the distance between cluster centroids. It provides an indication of how well-separated the clusters are. Higher intercluster distances suggest that the clusters are more distinct from each other. The Intercluster Distance D can be calculated as follows:

$$D = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^k d(c_i, c_j) \quad (1.16)$$

where:

- $d(c_i, c_j)$ is the distance between the centroids of clusters i and j .
- k is the number of clusters.

1.2.2.5 Intracluster Distance

Intracluster Distance [28] measures the average distance between points within the same cluster. Lower intracluster distances indicate that the points within a cluster are closely packed together, suggesting more cohesive clusters. The Intracluster Distance D_{intra} can be calculated as follows:

$$D_{\text{intra}} = \frac{1}{N} \sum_{i=1}^k \sum_{\substack{x, y \in C_i \\ x \neq y}} d(x, y) \quad (1.17)$$

where:

- N is the total number of data points.
- k is the number of clusters.
- C_i represents the i -th cluster.
- $d(x, y)$ is the distance between data points x and y .

1.2.2.6 Adjusted Rand Index

The Adjusted Rand Index [29] is a measure of the similarity between two data clusterings, adjusted for chance. It accounts for the fact that random clusterings are not expected to agree. ARI ranges from -1 to 1, where 1 indicates perfect agreement between the clusterings, 0 indicates random labeling, and negative values indicate worse-than-random agreement. The Adjusted Rand Index (ARI) is defined as:

$$ARI = \frac{RI - \text{Expected RI}}{\max(RI) - \text{Expected RI}} \quad (1.18)$$

where:

- RI (Rand Index) measures the percentage of correct decisions made by the algorithm.
- Expected RI is the expected value of RI for random clusterings.

1.2.2.7 Adjusted Mutual Information

The Adjusted Mutual Information [30] measures the agreement between two clusterings, adjusted for chance. It is based on the concept of mutual information

from information theory. AMI ranges from 0 to 1, where 1 indicates perfect agreement. The AMI is defined as:

$$AMI = \frac{MI(U, V) - E[MI(U, V)]}{\max(H(U), H(V)) - E[MI(U, V)]} \quad (1.19)$$

where:

- $MI(U, V)$ is the mutual information between clusterings U and V .
- $H(U)$ and $H(V)$ are the entropies of clusterings U and V , respectively.
- $E[MI(U, V)]$ is the expected mutual information between random clusterings.

1.2.2.8 Entropy

Entropy [31] is a fundamental concept in information theory and statistics, widely used in various fields including data analysis and clustering. In the context of clustering, entropy is a measure of uncertainty or randomness within the distribution of cluster assignments. It quantifies the degree of disorder or heterogeneity within the clustering result, providing insights into the cohesion and purity of the clusters.

The formula to compute entropy is given by:

$$\text{Entropy} = - \sum_{i=1}^k p_i \log_2(p_i) \quad (1.20)$$

where k is the number of clusters, and p_i represents the probability of occurrence of the i -th cluster label. This formula calculates the weighted average of the information content of each cluster label, where lower entropy values indicate more homogeneous clusters.

In clustering tasks, minimizing entropy helps in achieving well-separated and more cohesive clusters, leading to better-defined patterns and structures within the data.

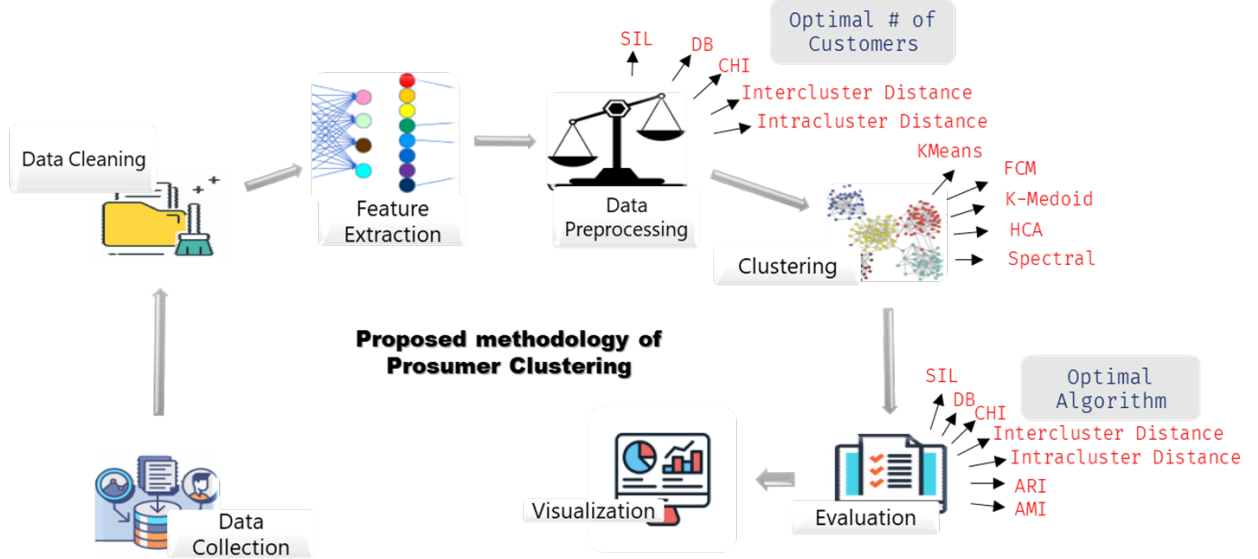


Figure 1.1: Proposed Methodology for Prosumer Clustering

1.3 Clustering Methodology

The proposed methodology for Prosumer Clustering on the datasets is shown in 1.1:

1.3.1 Data Cleaning

Data cleaning involved removing all unnecessary columns, focusing solely on consumption and generation data. Additionally, rows with missing or non-actual values were identified and excluded from the dataset. This step was crucial to ensure the integrity of the analysis, as incomplete or inaccurate data could significantly distort the clustering results.

After cleaning, the dataset was refined to contain a subset of customers, each with a complete set of data for the entire period. This cleaned dataset formed the basis for the subsequent analysis, allowing for a detailed examination of energy consumption and generation patterns among different customer groups.

1.3.2 Feature Extraction

Feature extraction is a crucial step that transforms raw data into a structured format suitable for clustering. Given the half-hourly granularity of the dataset, we derived meaningful statistics to reduce dimensionality while retaining critical information about consumption and generation patterns. The following steps were undertaken:

1. **Division into Time Windows:** The day was divided into four 6-hour

windows for General Consumption (GC) and Controlled Load Consumption (CL), and a single window for Gross Generation (GG) from 6:00 AM to 6:00 PM.

2. **Calculation of Mean and Peak Values:** For each window, mean (M) and peak (P) values were calculated.

This transformation facilitated the clustering analysis by converting half-hourly data into daily summary statistics.

1.3.3 Normalization

Normalization ensures that all features contribute equally to the clustering process by placing them on a comparable scale. This process is essential because features with larger magnitudes may dominate the distance calculations, leading to biased clustering results. By scaling the features, we ensure that each feature's contribution to the clustering process is proportional to its importance, thus preventing any single feature from unduly influencing the clustering outcome. We employed a normalization technique to transform the dataset, enhancing the accuracy and performance of the clustering algorithms. This step was critical in facilitating the identification of meaningful patterns during the clustering analysis.

1.3.4 Clustering Algorithm Selection

To determine the most effective methods for grouping the prosumer data, we evaluated a variety of clustering algorithms.

Each algorithm was applied to the normalized dataset, and their performance was assessed using multiple evaluation metrics.

1.3.4.1 Finding optimal number of clusters

To determine the optimal number of clusters, a set of algorithms including K-Means, Fuzzy C-Means (FCM), K-Medoids, Hierarchical Clustering, and Spectral Clustering were employed. These algorithms were selected based on their suitability for various data scenarios, as discussed in Section 1.1.2.1.

In assessing the clustering solutions provided by these algorithms, a suite of evaluation metrics was utilized. Specifically, the Silhouette Score , Calinski-Harabasz Index , Davies-Bouldin Index , Intercluster Distance , and Intraclass Distance were employed to gauge the quality and appropriateness of the clustering configurations generated by each algorithm. These evaluation metrics were elaborated upon in detail in Section 1.2.2.

1.3.4.2 Identifying best performing algorithm

After determining the optimal number of clusters for our dataset, we evaluated five clustering algorithms: K-Means, Fuzzy C-Means, Hierarchical Clustering, K-Medoids, and Spectral Clustering.

We utilized a set of evaluation metrics (discussed in 1.2.2) to assess their performance. These metrics collectively evaluate the compactness, separation, stability, and interpretability of the clusters. The performance of each algorithm was compared based on these evaluation metrics.

After analyzing the results, we identified the most suitable clustering algorithms for our dataset. These algorithms showed the best balance between cluster characteristics, making them optimal choices for further analysis and application in our research.

1.3.5 Visualization

After identifying the optimal clusters, various visualization techniques, including line graphs, barplots, scatter plots, and histograms, were utilized to examine the characteristics of each cluster. These visualizations offer insights into energy consumption and generation behaviors, facilitating informed decision-making and resource allocation strategies.

CHAPTER 2

Literature Survey

In recent years, the evolution of energy systems towards more decentralized and sustainable models has led to the emergence of prosumers—entities that both consume and produce energy. This transformation has sparked considerable interest in prosumer clustering within the context of energy management and smart grids. However, despite its potential to enhance demand response programs and improve energy efficiency, research on prosumer clustering remains limited.

The existing body of literature on prosumer clustering encompasses various studies that explore different clustering techniques for managing prosumers. While some papers focus on traditional clustering methods applied to prosumer data, others delve into specialized approaches tailored to time series clustering. Additionally, a subset of studies examines the intersection of machine learning and clustering techniques for optimizing energy management in smart grids.

In this literature review, we synthesize and analyze a comprehensive selection of papers related to prosumer clustering, time series clustering, and general clustering methods. By examining the methodologies, findings, and implications of these studies, we aim to provide insights into the current state of research, identify gaps in the literature, and highlight avenues for future investigation.

2.1 Prosumer Clustering in Smart Grids

Prosumer clustering plays a pivotal role in the optimization of energy management within smart grids, facilitating efficient demand response programs and promoting sustainable energy practices. In this section, we delve into key studies that have contributed to the understanding and implementation of prosumer clustering methodologies.

Silva et al. [32] focus on the application of partitional clustering methods to enhance energy management within smart grids. They emphasize the importance of

clustering techniques in aggregating prosumers based on their energy consumption and generation profiles. By comparing various partitional clustering algorithms, such as K-means and hierarchical clustering, the authors assess their effectiveness in grouping prosumers with similar energy patterns. Through empirical analysis, Silva et al. demonstrate the potential benefits of clustering in improving the reliability and efficiency of energy networks. The study provides valuable insights into the practical applications of clustering methods in smart grid environments, laying the groundwork for further research in this field.

Vergados et al. [33] explore the concept of virtual microgrids and their role in reducing energy costs through prosumer clustering. The study investigates the orchestration of prosumers into virtual microgrids using clustering algorithms such as spectral, genetic, and adaptive methods. By analyzing a real dataset of prosumers, the authors demonstrate how different clustering techniques can effectively group prosumers with similar energy profiles. Through the reduction of forecasting inaccuracies and the precise prediction of energy production and consumption patterns, Vergados et al. illustrate significant cost reductions achievable through clustering-based energy management strategies. The study underscores the financial benefits of prosumer clustering in energy markets and highlights the importance of accurate energy profiling in optimizing resource management.

Han et al. [34] propose the use of K-means clustering to improve the scalability of cooperative game theory-based energy sharing models among prosumers. By clustering energy profiles, their method enhances the computational efficiency of cooperative schemes while maintaining high financial incentives for prosumers. The study addresses the scalability challenges inherent in large-scale energy management systems by leveraging K-means clustering to group prosumers with similar energy characteristics. Through empirical evaluation, Han et al. demonstrate the practicality of K-means clustering in enhancing the scalability and efficiency of energy sharing models. The study highlights the significance of clustering techniques in addressing computational complexities and optimizing resource allocation in smart grids.

These studies collectively underscore the critical role of prosumer clustering in optimizing energy management within smart grids. By leveraging clustering algorithms and techniques, researchers aim to enhance the reliability, efficiency, and sustainability of energy systems, paving the way for a more decentralized and resilient energy future.

2.2 Management of Prosumers in Smart Grids

In the management of prosumers within smart grids, various approaches have been proposed to enhance energy efficiency and promote sustainable practices. This section explores key studies focusing on the management of prosumers through clustering techniques and demand response programs.

Rathnayaka et al. [35] introduce the concept of goal-oriented virtual prosumer communities to enhance the sustainability of smart grid energy sharing. The study emphasizes the formation, growth, and management of prosumer communities using clustering techniques to facilitate efficient energy auctioning and increased bargaining power in energy markets. By leveraging clustering algorithms, Rathnayaka et al. aim to create cohesive prosumer communities with aligned energy goals, fostering collaboration and resource optimization within smart grids.

Pelekis et al. [36] apply clustering techniques to design demand response (DR) programs for commercial and residential prosumers. Their study explores the implementation of K-means, K-medoids, and agglomerative clustering to create flexibility clusters, aiming to minimize reverse power flow and shift peak demand. Through the targeted design of DR programs based on clustering analysis, Pelekis et al. demonstrate the effectiveness of clustering algorithms in generating meaningful clusters that support tailored energy management strategies for prosumers.

Mythili and Madhiya [37] conduct a comprehensive comparative analysis of clustering algorithms in the context of data mining. The study evaluates the performance of various clustering techniques, including K-means, hierarchical clustering, and density-based clustering, in terms of their accuracy and computational efficiency. By comparing the advantages and shortcomings of different algorithms, Mythili and Madhiya provide insights into the selection of appropriate clustering methods for managing prosumers in smart grids. Their analysis contributes to the understanding of clustering algorithms' impact on energy management and informs the design of efficient clustering-based solutions for prosumer management.

These studies highlight the diverse approaches to managing prosumers within smart grids, emphasizing the role of clustering techniques in optimizing energy sharing, facilitating demand response programs, and enhancing resource allocation. By leveraging clustering algorithms, researchers aim to develop tailored solutions that address the unique challenges and opportunities presented by prosumer-centric energy systems.

2.3 Time Series Clustering Techniques

Time series clustering techniques play a crucial role in analyzing and managing temporal data, especially in the context of energy consumption and production within smart grids. This section explores key studies focusing on time series clustering methodologies and their applications.

Bindra and Mishra [38] present a detailed study of clustering algorithms, with a specific focus on their application to time series data. The study comprehensively analyzes the efficiency, advantages, and disadvantages of various clustering techniques in the context of time series analysis. By examining the performance of algorithms such as K-means, hierarchical clustering, and DBSCAN, Bindra and Mishra provide valuable insights into the suitability of different clustering methods for time series data clustering tasks.

Rodriguez et al. [39] systematically compare nine well-known clustering methods, emphasizing their performance in clustering time series data. The study evaluates algorithms such as K-means, DBSCAN, and spectral clustering on artificial datasets with varying properties, providing insights into the sensitivity of clustering methods to parameter configurations. Through their comparative analysis, Rodriguez et al. offer guidance on selecting appropriate clustering algorithms for time series data, highlighting the strengths and weaknesses of each method.

Gupta et al. [40] discuss the concept-based clustering approach for summarizing and understanding time series data. Their study explores the application of clustering techniques in various domains, such as gene and protein functionality analysis. By leveraging concept-based clustering, Gupta et al. demonstrate the utility of clustering methods in data compression and pattern recognition tasks. Their work underscores the versatility of clustering techniques in handling time series data and offers insights into innovative approaches for time series analysis.

These studies collectively contribute to the understanding of time series clustering techniques and their applications in energy management within smart grids. By exploring the efficiency and effectiveness of clustering algorithms, researchers aim to develop robust solutions for analyzing and managing temporal energy data, thereby enhancing the reliability and efficiency of smart grid systems.

2.4 Applications of Clustering in Smart Grids

Clustering algorithms find diverse applications in smart grids, ranging from consumer segmentation to demand response management. This section explores key studies focusing on the practical applications of clustering techniques in smart grid environments.

Xu and Tian [41] provide an extensive survey of clustering algorithms, analyzing their applications in various domains including smart grids. The study comprehensively discusses the strengths and weaknesses of different clustering techniques, offering insights into their suitability for different smart grid applications. By examining clustering algorithms from a broad perspective, Xu and Tian contribute to the understanding of clustering methods' role in modern science research and their relevance to smart grid technologies.

Ezugwu et al. [42] present a review of traditional and state-of-the-art clustering techniques, focusing on their applications in machine learning, data mining, and artificial intelligence. The study discusses the challenges associated with clustering high-dimensional datasets and uncertain cluster numbers, highlighting the need for improved clustering methods in smart grid applications. By providing a comprehensive overview of clustering techniques, Ezugwu et al. offer valuable insights into the potential advancements in clustering algorithms for smart grid management.

Mishra et al. [43] formulate subsequence clustering on residential building load data as a graph clustering problem, with applications in demand response management. The study introduces a novel quasi-clique-based graph clustering algorithm, which captures temporal dependencies and similarities between load subsequences. Through empirical evaluation, Mishra et al. demonstrate significant improvements over existing methods, emphasizing the importance of subsequence clustering in demand response management within smart grids.

These studies highlight the diverse applications of clustering techniques in smart grid environments, ranging from consumer segmentation to demand response management. By exploring the practical implications of clustering algorithms, researchers aim to develop efficient and scalable solutions for optimizing energy management and enhancing grid resilience.

2.5 Clustering Residential Electricity Customers

Clustering residential electricity customers is essential for effective demand management and personalized service provision. This section examines key studies that focus on clustering techniques for residential electricity customers.

Motlagh et al. [44] propose a strategy to cluster residential electricity customers by converting load time series into map models. Their method enhances cluster distinction and robustness against noise, leading to improvements in market segmentation and management. By leveraging clustering techniques, Motlagh et al. aim to develop tailored strategies for addressing the diverse needs of residential consumers in the electricity market.

Ruiz et al. [45] present a methodology for knowledge extraction in energy consumption data through time-series clustering techniques. The study compares several clustering methods to identify the most effective models for understanding energy consumption patterns. By applying clustering algorithms to energy consumption data, Ruiz et al. aim to uncover valuable insights that can inform decision-making and resource allocation in energy management systems.

Javed et al. [46] conduct a benchmark study on time series clustering using datasets from the UCR archive. Their comprehensive evaluation of clustering methods and distance measures provides valuable insights into the performance of different clustering techniques for residential electricity customer data. By benchmarking clustering algorithms, Javed et al. contribute to the understanding of the suitability and effectiveness of clustering methods in the context of energy consumption analysis.

These studies underscore the importance of clustering techniques in understanding and managing residential electricity consumption. By clustering customers based on their energy usage patterns, researchers aim to develop targeted strategies for demand management and service optimization in residential energy systems.

2.6 Time Series Clustering Methods

Time series clustering methods play a crucial role in analyzing sequential data and extracting meaningful patterns. This section explores key studies that focus on clustering techniques specifically tailored for time series data.

Özkoç [47] examines time series clustering methods, providing a detailed taxonomy and case study on gene expression data. By categorizing clustering algorithms based on data representation, similarity measures, and clustering techniques, Özkoç offers valuable insights into the application of clustering methods in various fields.

Alqahtani et al. [48] review deep time-series clustering (DTSC) methods, highlighting recent advancements and state-of-the-art techniques. Their study emphasizes the potential of deep learning approaches in clustering complex time series data, providing a roadmap for future research in this area.

Wang et al. [49] propose a scalable method for time series clustering using self-organizing maps and global measures of time series characteristics. Their approach addresses the challenges of high dimensionality and long time series, demonstrating robustness and efficiency in clustering large datasets.

These studies highlight the importance of time series clustering methods in analyzing sequential data and uncovering hidden patterns. By leveraging clus-

tering techniques tailored for time series data, researchers aim to improve the understanding and interpretation of complex temporal phenomena.

2.7 Machine Learning-Based Clustering

Machine learning-based clustering techniques offer advanced capabilities for analyzing electricity load profiles and optimizing energy management strategies. This section explores key studies that leverage machine learning approaches for clustering in smart grid environments.

Michalakopoulos et al. [50] present a machine learning-based framework for clustering residential electricity load profiles. By applying multiple clustering algorithms and leveraging Explainable AI, their study achieves optimal load profiling and enhances the interpretability of clustering results. The framework demonstrates practical applications of advanced clustering techniques in demand response programs.

Wang et al. [51] conduct an extensive review of smart meter data analytics, focusing on applications, methodologies, and challenges. They emphasize the importance of clustering techniques in consumer segmentation, load forecasting, and anomaly detection, illustrating how these methods can enhance the efficiency and sustainability of power grids.

These studies underscore the significance of machine learning-based clustering techniques in addressing complex challenges in smart grid management. By leveraging advanced algorithms and analytics, researchers aim to improve the reliability, efficiency, and sustainability of modern energy systems.

2.8 Probabilistic Load Flow Approaches

Probabilistic load flow approaches are essential for modeling uncertainties in power distribution systems, particularly with the integration of renewable energy sources and electric vehicle charging. This section examines a key study that focuses on probabilistic load flow approaches in smart grid environments.

Ramadhania et al. [52] review probabilistic load flow (PLF) approaches for power distribution systems with photovoltaic (PV) generation and electric vehicle (EV) charging. The study highlights the need for probabilistic methods to handle uncertainties in load and generation, emphasizing the role of clustering in modeling these uncertainties and improving PLF analyses.

The research underscores the importance of probabilistic load flow approaches in enhancing the reliability and efficiency of power distribution systems, particu-

larly in the context of increasing renewable energy integration and electric vehicle adoption.

2.9 Distributed Energy Resources Coordination

Coordination of distributed energy resources (DERs) is crucial for ensuring equitable and efficient utilization of renewable energy sources within smart grid environments. This section explores a study that addresses the fair coordination of distributed energy resources.

Gebbran et al. [53] develop a distributed optimal power flow (DOPF) method for fair coordination of distributed energy resources. Their approach ensures equitable PV curtailment and incorporates Volt-Var control, using clustering techniques to manage the distribution of resources and optimize power flow.

The study highlights the importance of fairness in DER coordination and underscores the role of clustering techniques in achieving equitable resource allocation and enhancing the efficiency of distributed energy systems.

2.10 Energy Sharing and Optimization

Effective energy sharing mechanisms are essential for optimizing resource utilization and promoting sustainable energy practices. This section discusses studies focused on energy sharing and optimization strategies within smart grid environments.

Liu et al. [54] introduce an energy-sharing provider (ESP) equipped with energy storage to facilitate sharing among PV prosumers. Their hybrid approach combines stochastic programming and Stackelberg game theory, leveraging clustering techniques to optimize day-ahead scheduling and real-time demand response.

Holtschulte et al. [55] explore the implementation of local energy markets (LEMs) within the Clustering Power System Approach (CPSA). They use genetic algorithms for economic optimization and clustering techniques to manage smart prosumer participation. Their results demonstrate the economic benefits of integrating LEMs with CPSA for smart grid operation.

These studies highlight the significance of energy sharing and optimization strategies in enhancing economic efficiency, promoting renewable energy integration, and fostering collaboration among prosumers within smart grid ecosystems.

2.11 Solar Prosumer Identification

Identifying solar prosumers is crucial for grid operators to manage the integration of distributed solar energy resources effectively. In this section, we discuss a study focused on the identification of solar prosumers within smart grid environments.

Donaldson and Jayaweera [56] propose a method for effectively identifying solar prosumers by analyzing net smart meter data. Their advanced classification approach utilizes agglomerative clustering to accurately detect and classify prosumers with solar photovoltaic (PV) installations. By improving the accuracy of prosumer identification, this method enhances the situational awareness of grid operators and aids in better circuit protection and voltage regulation settings.

This study contributes to the optimization of grid operations by providing a robust solution for managing the growing number of solar prosumers and facilitating the seamless integration of distributed solar energy resources into the grid.

Prosumer Clustering with Ausgrid Dataset

In our exploration of prosumer behavior within the energy market, we utilized the Ausgrid dataset to gain insights into residential electricity consumption and solar generation patterns. This chapter presents the dataset used, the preprocessing steps undertaken, and the methodologies applied, followed by a comprehensive discussion of the results and their implications for energy management strategies.

3.1 Dataset Overview

The Ausgrid dataset [57], provided by Ausgrid, a major electricity distributor in Australia, comprises comprehensive data on residential electricity consumption and generation from solar photovoltaic systems. This dataset is invaluable for researchers, policymakers, and industry stakeholders aiming to analyze and optimize energy consumption patterns, enhance demand response programs, and improve grid management.

The dataset includes data from 300 randomly selected solar customers within the Ausgrid network, collected over three years, from 1 July 2010 to 30 June 2013. These customers were billed on a domestic tariff and had a gross metered solar system installed throughout the period. The dataset includes:

While the dataset offers rich insights, it is important to acknowledge certain limitations. The sample of 300 customers, though randomly selected, may not fully represent the broader population of residential solar customers in the Ausgrid network. Additionally, data gaps and estimation errors, though minimized through cleaning, may still impact the analysis.

In summary, the Ausgrid dataset provides a robust foundation for examining prosumer behavior through advanced clustering techniques. The following sections

Column Name	Description
Customer ID	A unique identifier for each customer.
Postcode	The location identifier, indicating the geographical area of each customer.
Generator Capacity	The capacity of the solar panels, measured in kilowatt peak(kWp).
Consumption Category	The type of consumption data, categorized into General Consumption(GC), Controlled Load Consumption (CL), and Gross Generation(GG).
Date	Recorded in DDMMYYYY format.
Half-hour Interval Data	Columns from 0:00 to 23:30 representing kilowatt hours(kWh) of electricity consumed or generated in each half-hour interval.
Row Quality	An indicator of data quality, specifying whether the data is actual or estimated.

Table 3.1: Description of the columns in the Ausgrid dataset.

detail the methodology employed for this analysis, including feature extraction, normalization, and clustering, followed by a discussion of the results.

3.2 Cluster Analysis

This section presents the results of clustering analysis performed on the Ausgrid dataset, aiming to identify distinct groups of prosumers based on their energy usage behaviors.

The proposed clustering methodology discussed in Section 1.3 is applied to the Ausgrid dataset.

First, data cleaning, feature extraction into various time windows, and pre-processing was done to make the data suitable for clustering. For clustering, the optimal number of clusters is unknown beforehand and difficult to predict. Therefore, various evaluation scores like Silhouette Score (SIL), Calinski-Harabasz Index (CHI), Davies-Bouldin Index (DBI), Intercluster Distance, and Intraccluster Distance mentioned in detail in Section 1.2.2 were evaluated for a range of 2 to 11, and the optimal value k (number of clusters) was found to be 3. Figures 3.1, 3.3, and 3.2 show the Silhouette Scores, Calinski-Harabasz Index, and Davies-Bouldin Index for various cluster counts, respectively.

Moving on to the next part of the methodology, the best algorithm is chosen for $k = 3$. This is done by checking the SIL, CHI, DBI, Intercluster and Intraccluster Distances, Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), and Entropy shown in 3.2 for all five algorithms explained in detail in 1.3.4. Table 3.2

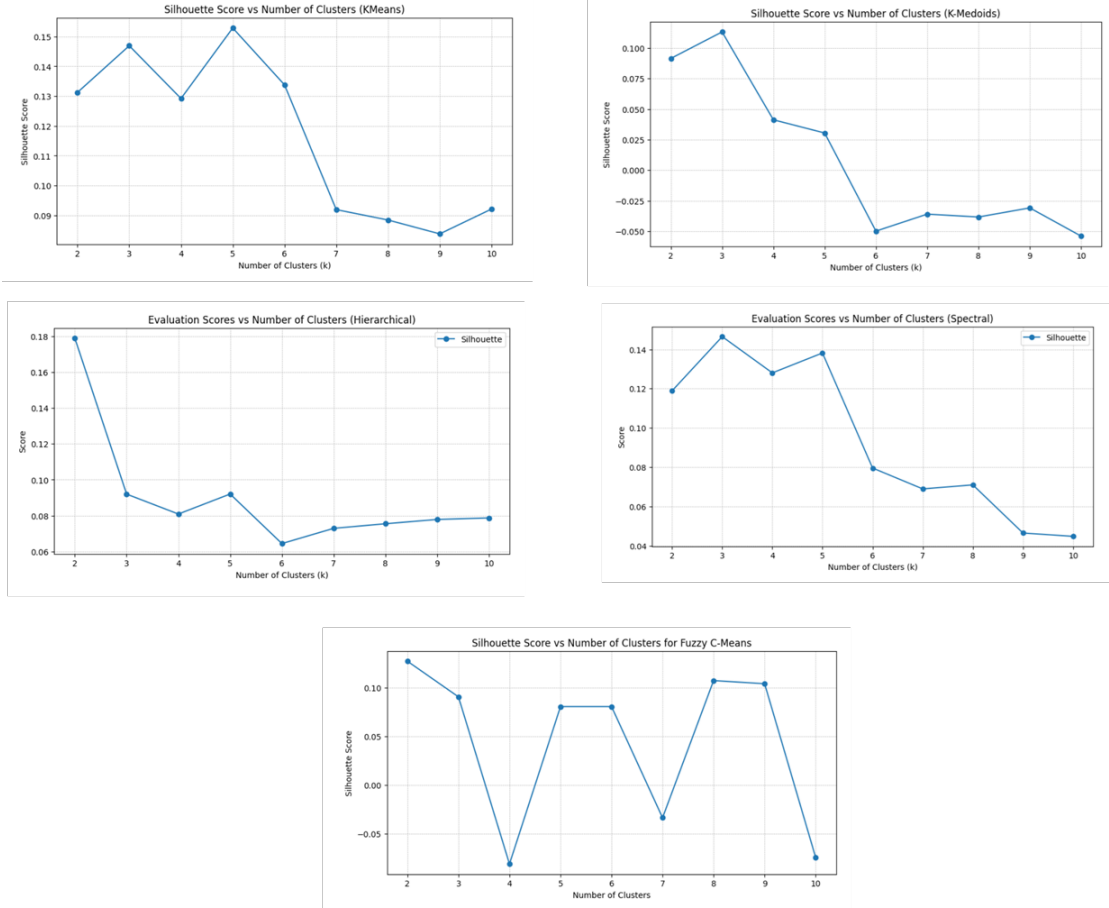


Figure 3.1: Silhouette Score Comparison

shows the scores for all the algorithms.

 Table 3.2: Evaluation Scores and Distances for Clustering Algorithms ($k = 3$)

Algorithm	SIL	DB	CHI	Intracuster	Intercluster	ARI	AMI	Entropy
K-Means	0.146925	2.062306	19.515423	24.926171	17.299196	-0.009188	-0.009324	1.523795
Fuzzy C-Means	0.130413	2.458186	19.188087	26.441574	10.745621	-0.002329	-0.002252	0.999955
Hierarchical	0.091985	2.242535	15.909189	25.445320	16.885849	0.005092	0.002881	1.456524
K-Medoids	0.113282	2.320753	16.721735	25.314752	15.637266	-0.010297	-0.009727	1.569826
Spectral	0.146476	2.078459	19.233029	24.960926	17.108929	-0.008330	-0.009173	1.530183

The evaluation metrics indicated that K-Means was the best-performing algorithm for this dataset, providing well-defined and meaningful clustering configurations. However, there is room for improvement in the clustering results, particularly in clusters where the intra-cluster distances were relatively high, indicating lower cohesion.

Next the best performing algorithm, that is, KMeans was applied on the normalized data and clusters were obtained. The clustering results revealed distinct patterns in the energy consumption behaviors of different prosumer groups.

The time-series patterns for general consumption (Figure 3.5) provide insights into the consumption trends within each cluster, showing that Cluster 1 peaks

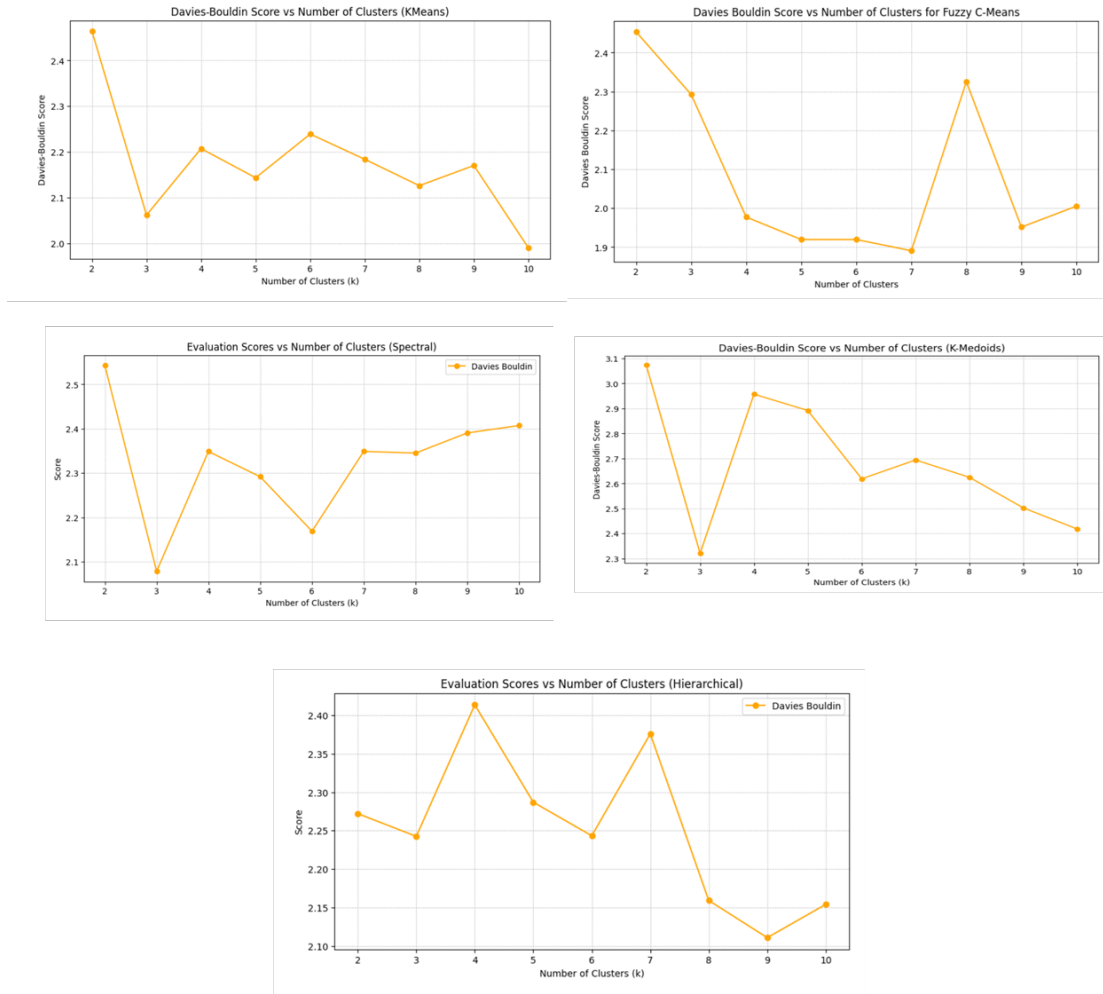


Figure 3.2: Davies-Bouldin Index Comparison

in the evening, Cluster 2 has flexible usage, and Cluster 3 maintains consistent low usage. The time-series patterns for gross generation (Figure 3.4) indicate similar generation trends across the clusters, with Cluster 1 having the highest average generation. The time-series patterns for controlled load (Figure 3.6) reveal different peak times and usage patterns for each cluster, indicating diverse energy management strategies.

Next, cluster visualization was performed. A PCA dimensionality reduction was also used for visualization. The PCA scatter plot (Figure 3.7) illustrates the clear separation between the clusters with minimal overlapping, indicating distinct energy usage behaviors.

The surplus/deficit analysis (Figure 3.8) shows that Cluster 1 and Cluster 3 have the highest surplus, meaning the amount left after consuming electricity after generating is high. This indicates that members of these clusters can be actively targeted for demand-side management programs where prosumers can give their surplus energy to the utility company and receive compensation.

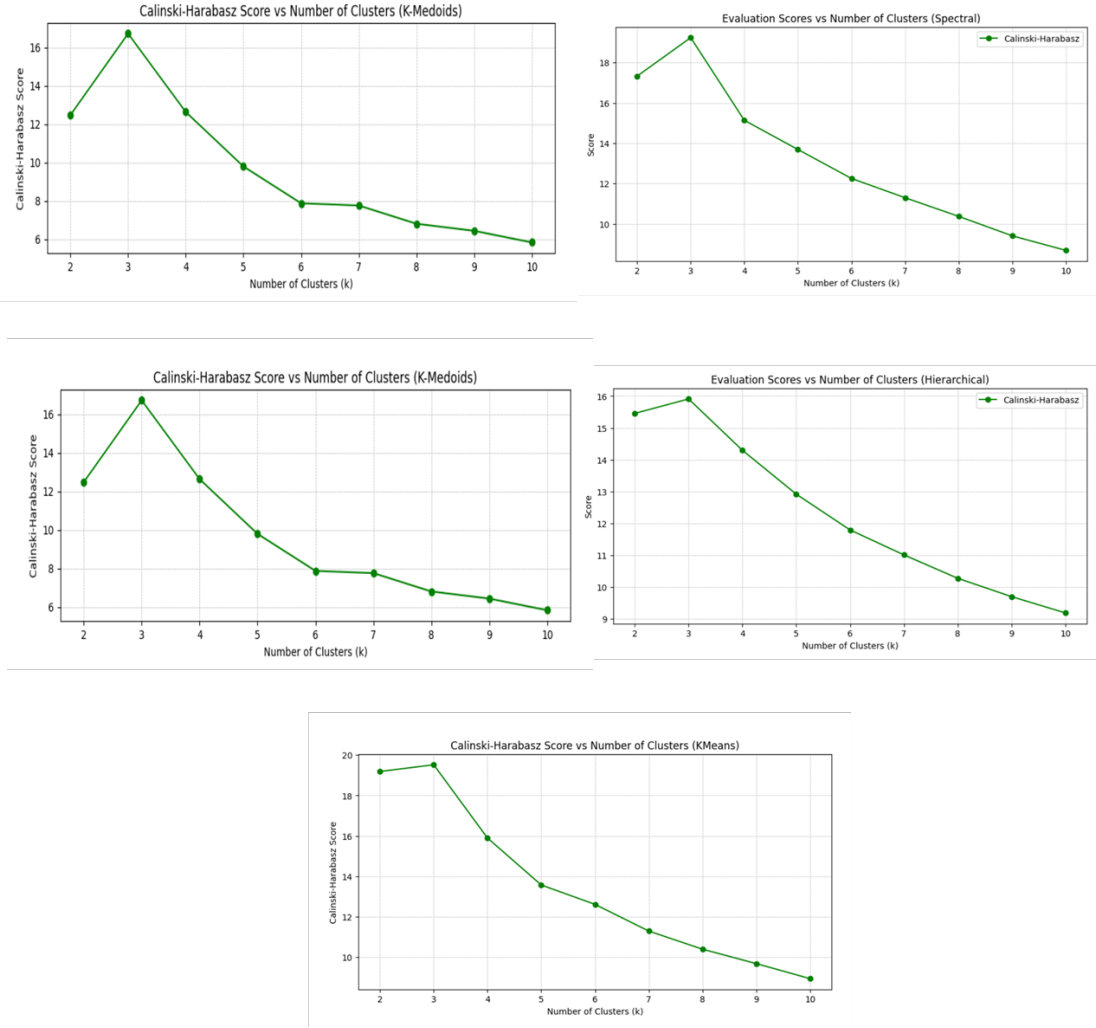


Figure 3.3: Calinski-Harabasz Index Comparison

The cluster analysis of the Ausgrid dataset successfully identified three distinct groups of prosumers with varying energy consumption patterns. These clusters provide valuable insights for tailoring DR programs to specific customer segments. For example, households with high evening usage may benefit from Time-of-Use (ToU) pricing, while those with flexible usage patterns could be more suitable for Critical Peak Pricing (CPP) programs.

3.3 Interpretation of Gross Generation Capacity

In this section, we focus on the analysis of Gross Generation (GG) data to interpret the generation capacity of prosumers. For this study, only the GG data was considered without any feature extraction, using data from a single year. The clustering methodology discussed previously was applied, and the optimal number of clusters was determined to be 3, with K-Means being the chosen algorithm

again.

3.3.0.1 Clustering Results

The clustering results revealed distinct patterns in the GG behaviors of different prosumer groups.

Figure 3.9 shows the Sum of Squared Distances (SSD) vs the number of clusters (k) for the GG data. The optimal number of clusters was determined to be 3 based on the elbow method.

Table 4.1 summarizes the evaluation scores and distances for different clustering algorithms when $k = 3$:

Table 3.3: Evaluation Scores and Distances for Clustering Algorithms ($k = 3$)

Algorithm	SIL	DB	CHI	Intracuster	Intercluster	ARI	AMI	Entropy
K-Means	0.398184	1.003665	59028.124656	1.353010	2.534123	-0.000041	-0.000922	1.561649
FCM	0.398122	1.015134	58984.464721	1.357647	2.549408	0.000007	0.000006	1.556766
Hierarchical	0.349490	1.002673	49719.043038	1.414698	2.623744	0.000041	0.000039	1.577180
K-Medoids	0.402782	0.950941	58122.255334	1.339189	2.503940	0.000024	0.000018	1.569471
Spectral	0.269494	0.809795	23808.353547	1.651910	2.135526	0.000062	0.000043	1.299958

3.3.0.2 Visualization of Clusters

To better understand the clustering results, various visualizations were created, including the GG patterns of the three clusters, the average GG energy by cluster, a PCA plot, the peak GG for each member of each cluster, and a plot of GG showing the line graphs of all customers along with medians distinguishing each cluster.

3.3.0.3 Discussion on Visualizations

The PCA plot (Figure 3.10) illustrates the separation between the clusters, indicating distinct generation behaviors.

The GG patterns for each cluster (Figure 3.11) provide insights into the generation trends within each group.

The average GG energy bar plot (Figure 3.12) shows the typical generation capacity for each cluster, while the peak GG plot (Figure 3.13) highlights the maximum generation levels reached by prosumers in each cluster.

The GG line graphs with medians (Figure 3.14) help visualize the distribution and central tendencies within the clusters.

Figure 3.13 shows the GG peak values and corresponding GC values at that time for clusters 1, 2, and 3, for a sample of customers. From these figures, we can observe that the majority of members in clusters 2 and 3 have higher

GG peak values compared to their corresponding GC values at the same time. This indicates that these clusters can be targeted for demand-side management programs, allowing them to independently contribute surplus energy to the utility grid.

The cluster analysis of GG data identified three distinct groups of prosumers with varying generation patterns. These clusters provide valuable insights for optimizing energy generation and improving grid management strategies. For instance, clusters with higher average generation capacities can be targeted for programs that incentivize surplus energy contribution to the grid.

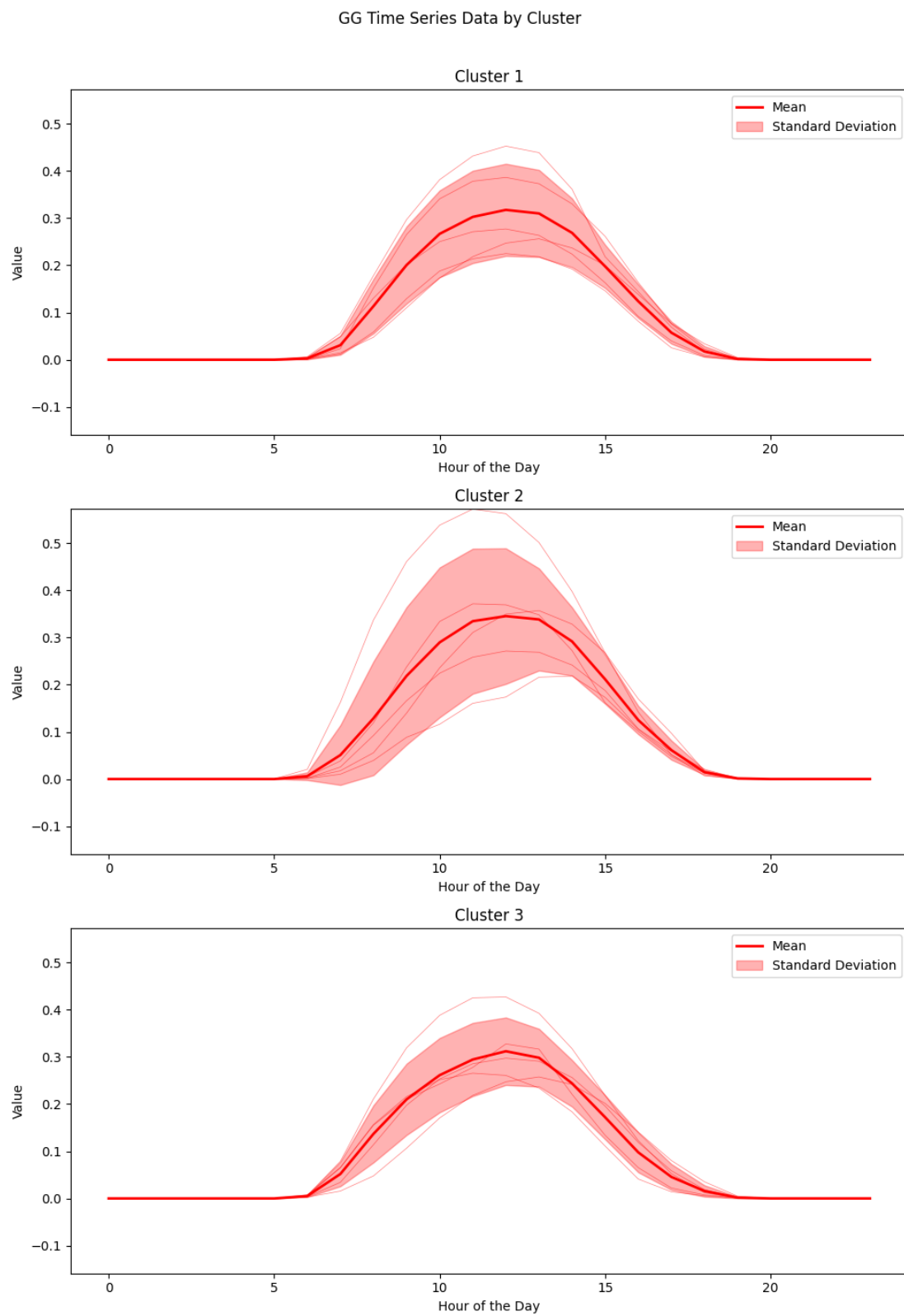


Figure 3.4: Time-Series Patterns of the Three Clusters for Gross Generation (GG)

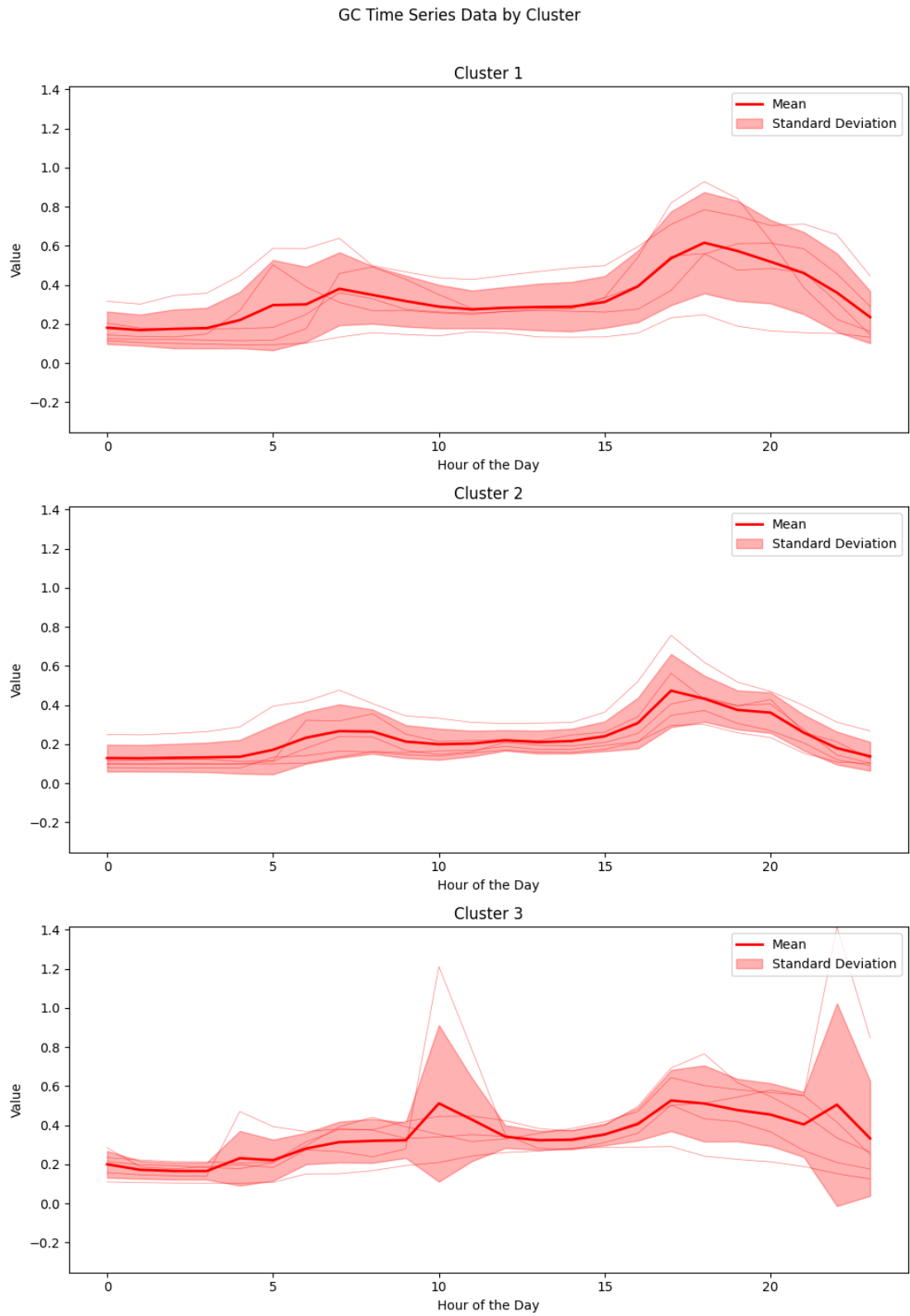


Figure 3.5: Time-Series Patterns of the Three Clusters for General Consumption (GC)

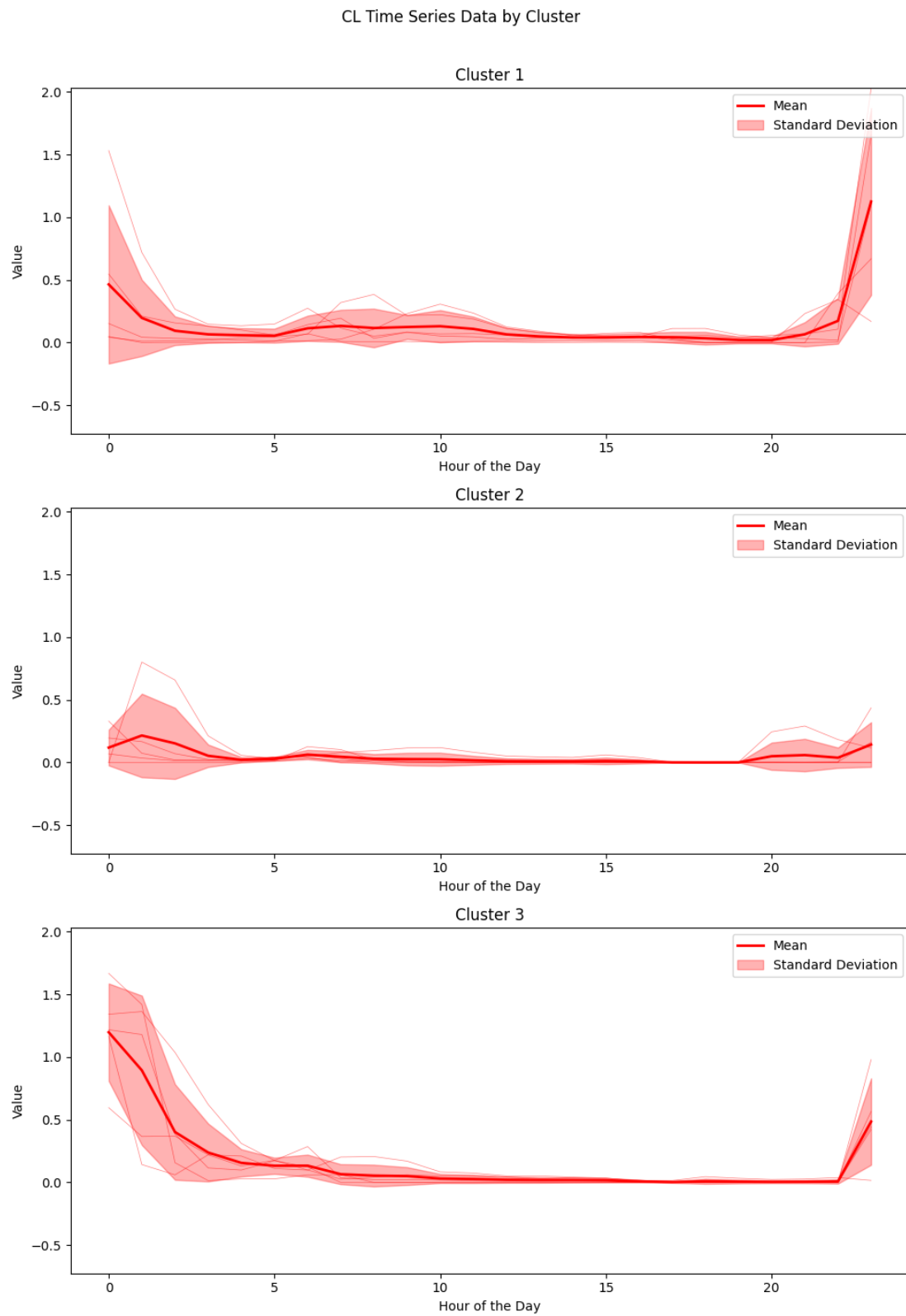


Figure 3.6: Time-Series Patterns of the Three Clusters for Controlled Load (CL)

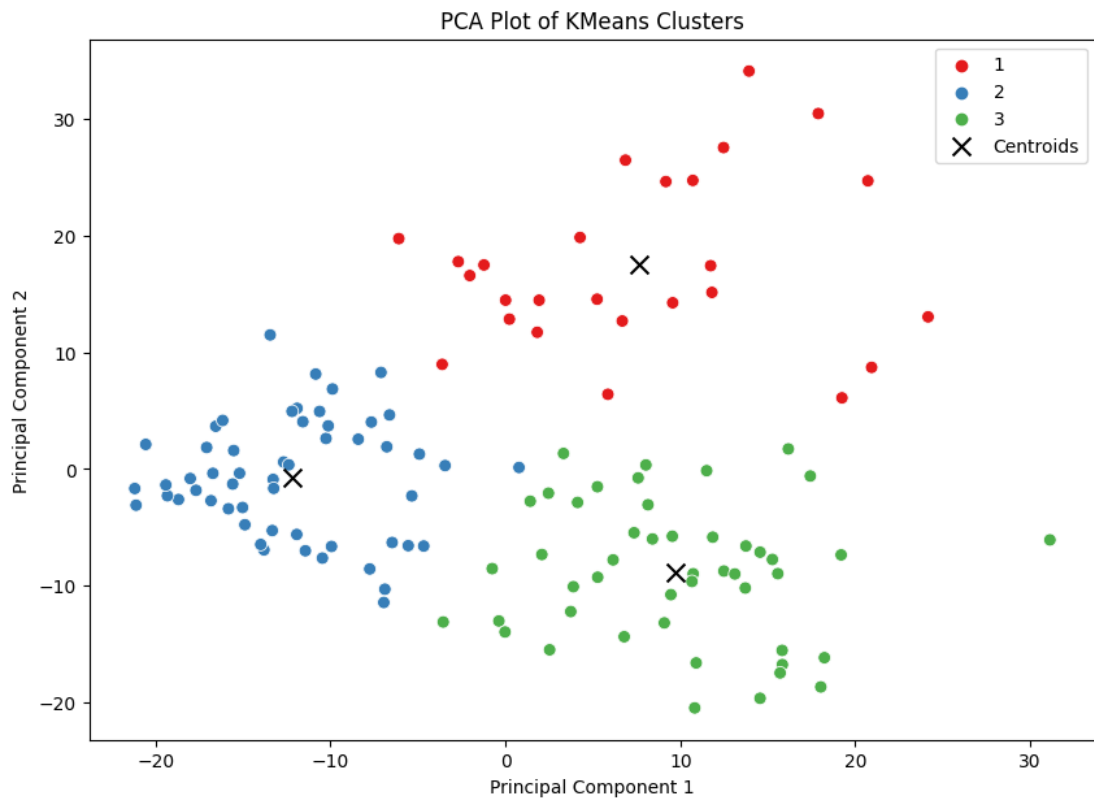


Figure 3.7: PCA Scatter Plot of KMeans Clusters

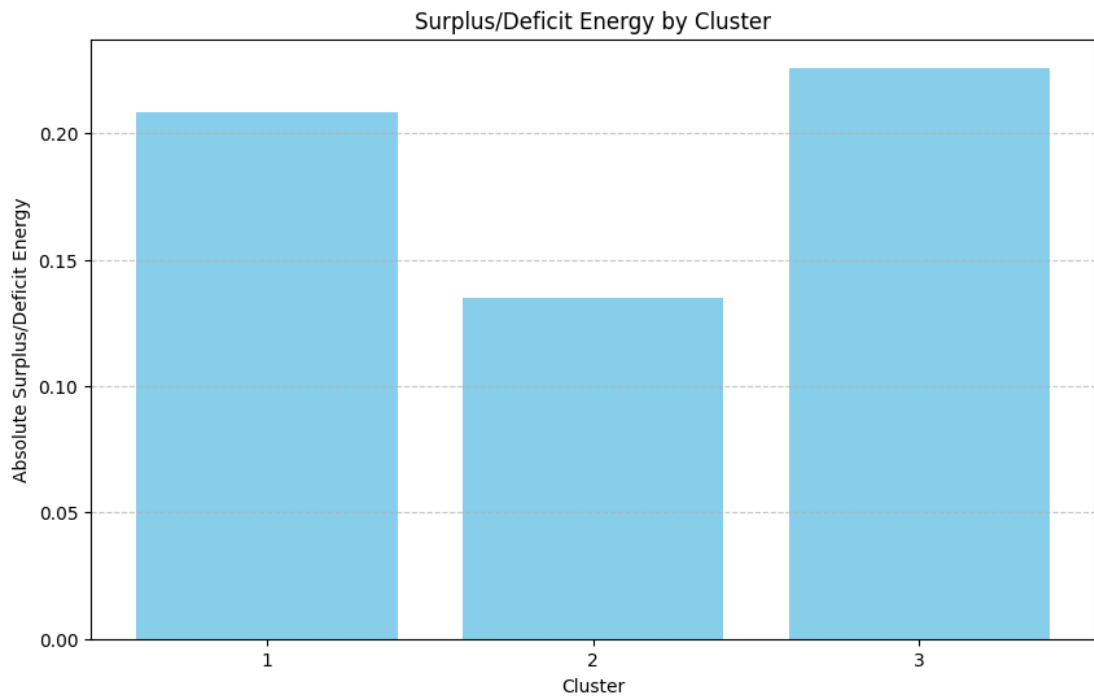


Figure 3.8: Surplus/Deficit Energy by Cluster

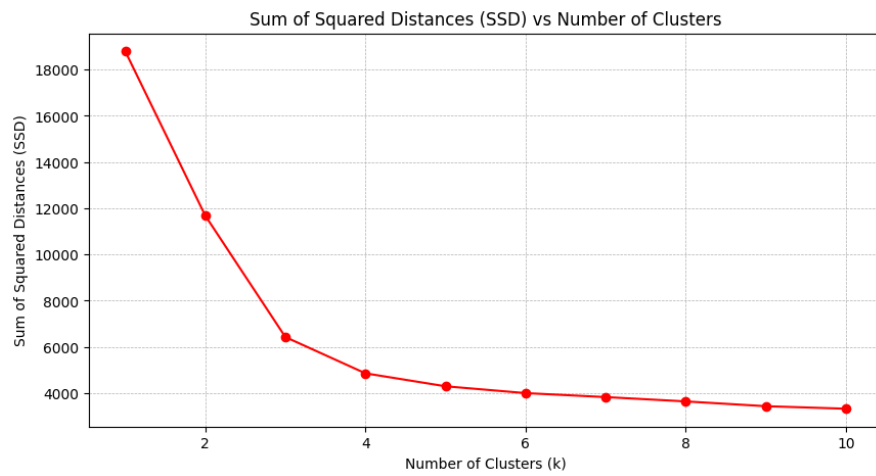


Figure 3.9: Sum of Squared Distances (SSD) vs Number of Clusters (k) for GG Data

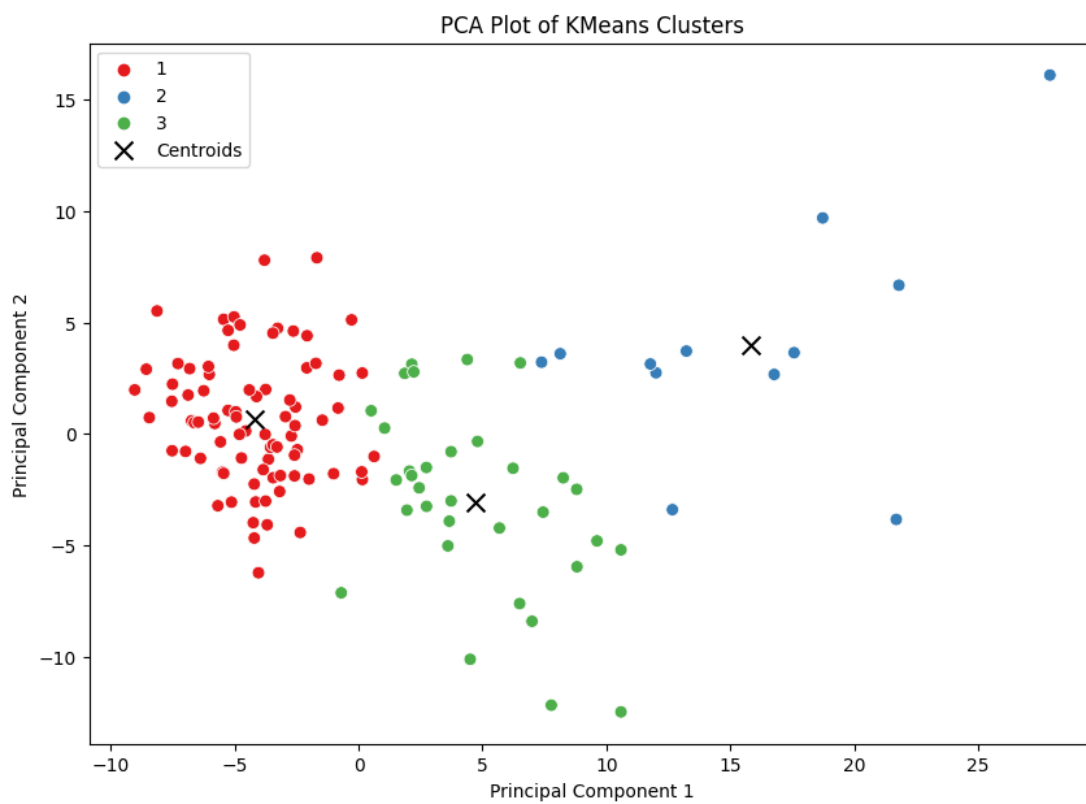


Figure 3.10: PCA Plot of GG Clusters

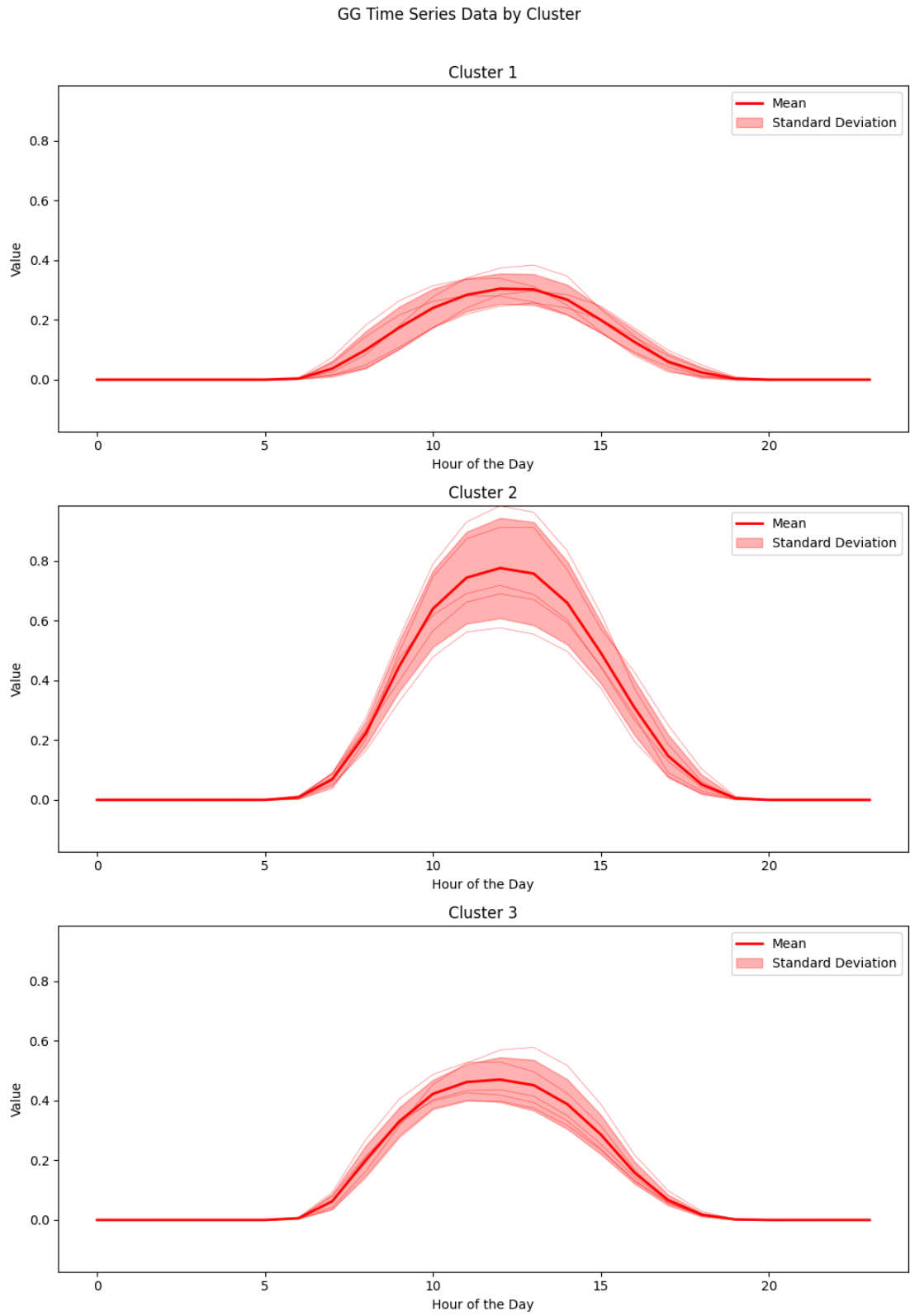


Figure 3.11: GG Patterns of the Three Clusters

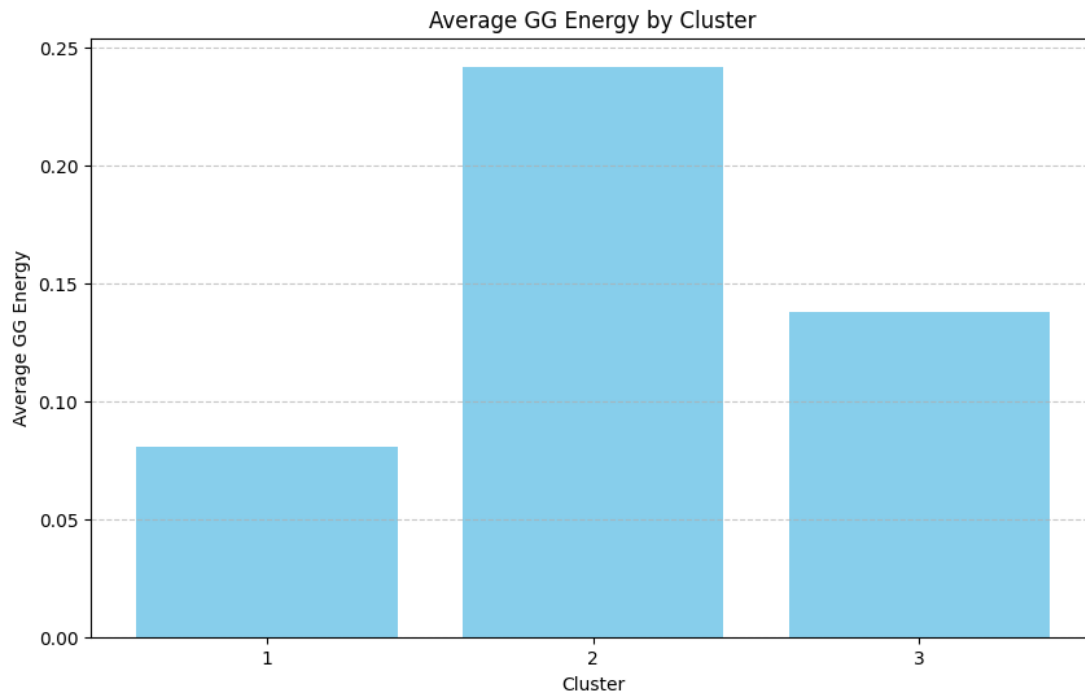


Figure 3.12: Average GG Energy by Cluster

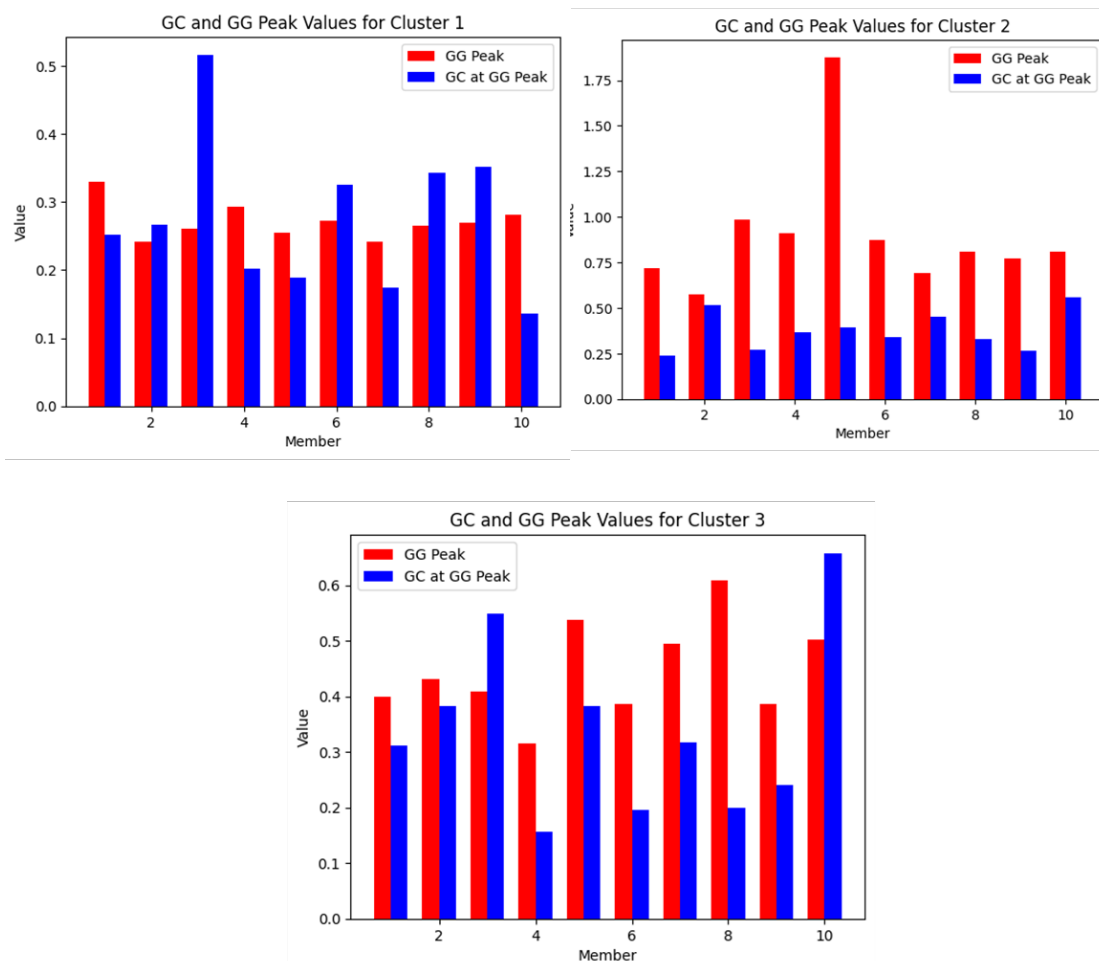


Figure 3.13: Peak GG for Each Member of Each Cluster

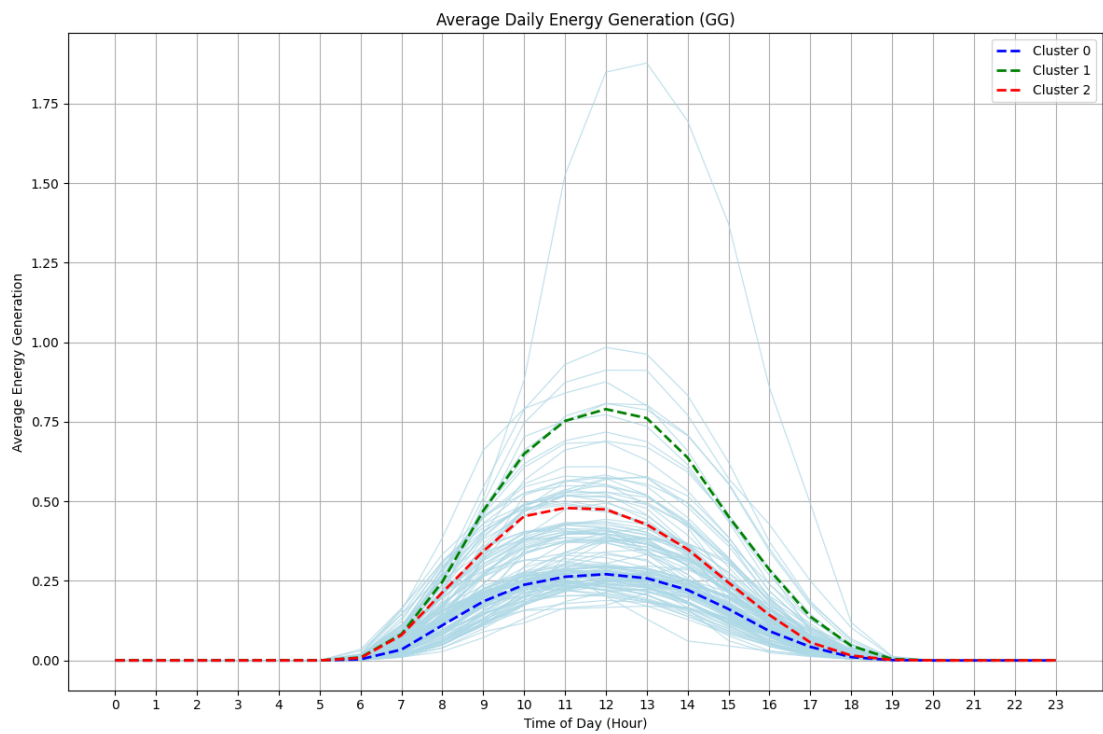


Figure 3.14: GG Line Graphs of All Customers with Medians Distinguishing Each Cluster

Prosumer Clustering with Synthetic Dataset

This chapter focuses on analyzing the Danish dataset, which is instrumental for research in smart grid management and demand response strategies. We begin with an overview of the dataset, followed by a description of the methodology used for sampling, data preparation, feature extraction, normalization, and clustering. The results section presents the patterns identified in energy consumption and production, as well as the distribution of seasons within the clusters. We conclude with a discussion on the implications for demand response programs and strategic recommendations based on our findings.

4.1 Dataset Overview

The Danish dataset [58], a synthetic compilation designed to emulate residential electricity consumption and production patterns in Denmark, serves as a valuable tool for research in smart grid management and demand response strategies. This dataset was generated to provide a comprehensive and controlled environment for analyzing energy behaviors under various conditions.

The dataset encompasses 600,000 days of energy data, derived from the consumption profiles of 2,000 real Danish consumers. It includes detailed records of imported energy from the grid and exported energy to the grid, segmented by photovoltaic (PV) systems. The dataset is available in two formats: a nested pickle file for use in Python and an XLSX file for users less familiar with computational tools.

Key Components

1. Energy Consumption and Generation Data

- **Imported Energy (GC):** Energy consumed from the grid.
- **Exported Energy (GG):** Energy generated by residential PV systems and exported back to the grid.

2. Temporal Segmentation

- **Seasonal Categorization:** Data classified by seasons (spring, summer, autumn, winter).

4.2 Clustering Analysis

Following the proposed clustering methodology mentioned in Section 1.3, we applied it to the Danish dataset. Each row in the dataset corresponds to a day, and we grouped the days accordingly. No specific dates are mentioned, and the rows are not in any particular sequence. To ensure a balanced representation of different seasons, the dataset was sampled such that each season had an equal number of records. This approach resulted in 11,995 records for each season, ensuring that the analysis could capture the effects of seasonal variations on energy consumption and generation patterns.

Table 4.1: Evaluation Scores and Distances for Clustering Algorithms ($k = 3$)

Algorithm	SIL	DB	CHI	Intraccluster	Intercluster	ARI	AMI	Entropy
K-Means	0.398184	1.003665	59028.124656	1.353010	2.534123	-0.000041	-0.000922	1.561649
Fuzzy C-Means	0.398122	1.015134	58984.464721	1.357647	2.549408	0.000007	0.000006	1.556766
Hierarchical	0.349490	1.002673	49719.043038	1.414698	2.623744	0.000041	0.000039	1.577180
K-Medoids	0.402782	0.950941	58122.255334	1.339189	2.503940	0.000024	0.000018	1.569471
Spectral	0.269494	0.809795	23808.353547	1.651910	2.135526	0.000062	0.000043	1.299958

The table (Table 4.1) presents the evaluation scores and distances for various clustering algorithms with $k = 3$. The results indicate that K-Medoids performs slightly better in terms of silhouette and Davies-Bouldin scores, suggesting well-defined and compact clusters. Spectral clustering shows a significantly lower Davies-Bouldin score, indicating good separation between clusters but at the cost of higher intraccluster distances.

Clustering Results

The clustering results 4.2 revealed distinct patterns in the GG and GC behaviors of different days.

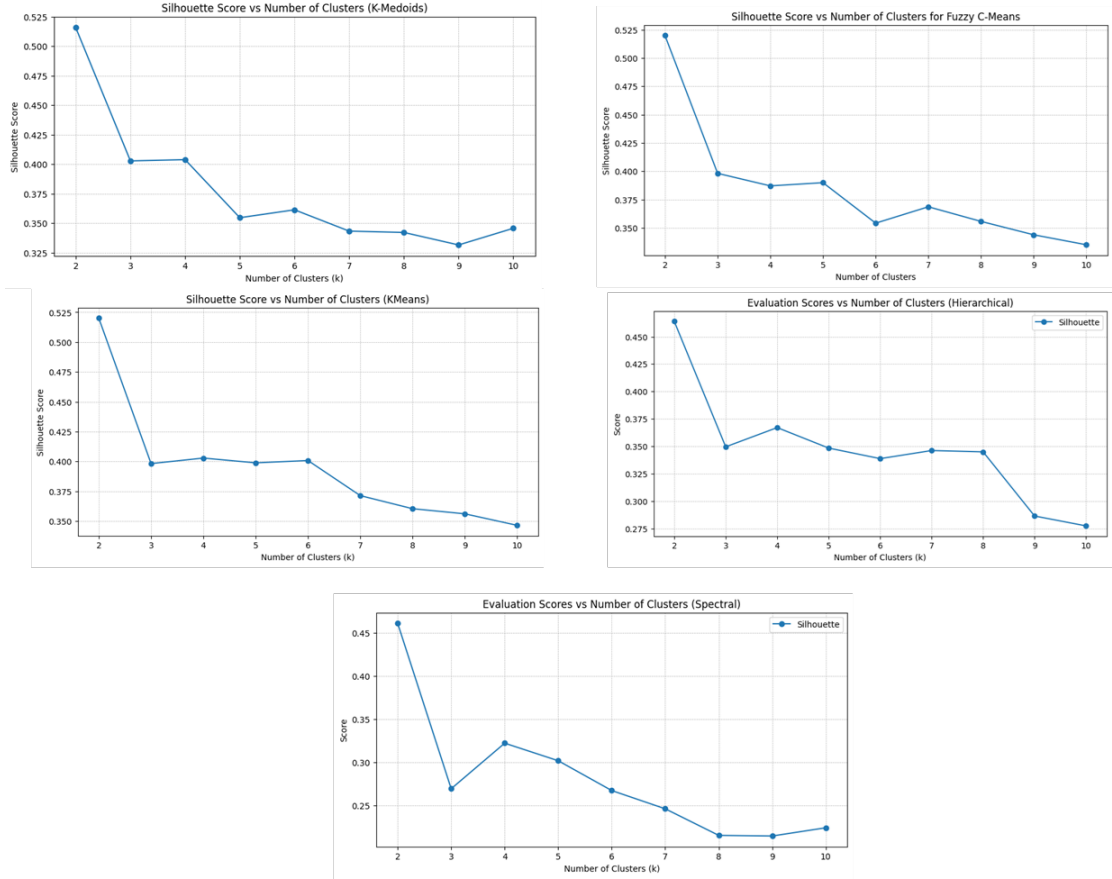


Figure 4.1: Silhouette Score vs Number of Clusters for Different Algorithms

Clustering Results

The clustering results revealed distinct patterns in the GG and GC behaviors of different days.

The GG time series data by cluster (Figure 4.2) shows the mean and standard deviation of GG values for each cluster across different hours of the day. Cluster 1 exhibits a peak during the early morning hours, suggesting high energy consumption at the start of the day. Cluster 2 shows a more distributed pattern of consumption throughout the day, while Cluster 3 indicates a significant peak during midday hours. This variation in patterns indicates different usage behaviors among the clusters, which can inform tailored demand response strategies.

Similarly, the GC time series data by cluster (Figure 4.3) presents the mean and standard deviation of GC values for each cluster throughout the day. Cluster 1 shows higher energy generation in the early hours, Cluster 2 exhibits a broader peak spanning midday, and Cluster 3 has a more pronounced peak in the late afternoon. These generation patterns provide insights into the optimal times for energy production, which can enhance scheduling for energy storage and utilization.

The evaluation of clustering algorithms using different metrics provides com-

prehensive insights. The silhouette score analysis (Figure 4.1) illustrates how the clustering performance varies with the number of clusters for different algorithms. Higher silhouette scores indicate better-defined clusters. The analysis suggests that K-Means and K-Medoids perform best with a smaller number of clusters (around 2-4), while the performance decreases with a higher number of clusters. This helps in selecting the optimal number of clusters for effective grouping.

The Davies-Bouldin score (Figure 4.4) reveals that Spectral and Hierarchical clustering algorithms achieve lower DB scores with a smaller number of clusters, indicating more compact and well-separated clusters.

The Calinski-Harabasz score (Figure 4.5) show that K-Means achieve higher CHI scores with fewer clusters, supporting the findings from the silhouette score and DB score analyses.

The PCA plot of KMeans clusters (Figure 4.6) provides a visual representation of the clusters in a reduced dimensional space. The clear separation of clusters along the principal components indicates distinct groupings, validating the effectiveness of the K-Means algorithm in identifying unique energy consumption and generation patterns.

Analysis of Seasonal Impact

The analysis of the day-wise distribution of clusters across different seasons and the count of each season in each cluster reveals important insights. As shown in Figures 4.7 and 4.8, there is no clear pattern that links the clustering outcomes strictly to seasonal variations. Each cluster contains days from all seasons, suggesting that factors other than seasonality play a significant role in defining energy consumption and generation patterns.

These findings indicate that it would be incorrect to attribute the variations in energy behavior solely to seasonal effects. Other factors, such as household characteristics, weather conditions, and consumer behavior, are likely to have a more substantial impact on energy patterns. Further investigation into these factors is recommended to develop a more comprehensive understanding of the determinants of energy usage and production.

By considering a broader range of influencing factors, future research can better inform demand response strategies and the design of targeted interventions for improving energy efficiency and grid management.

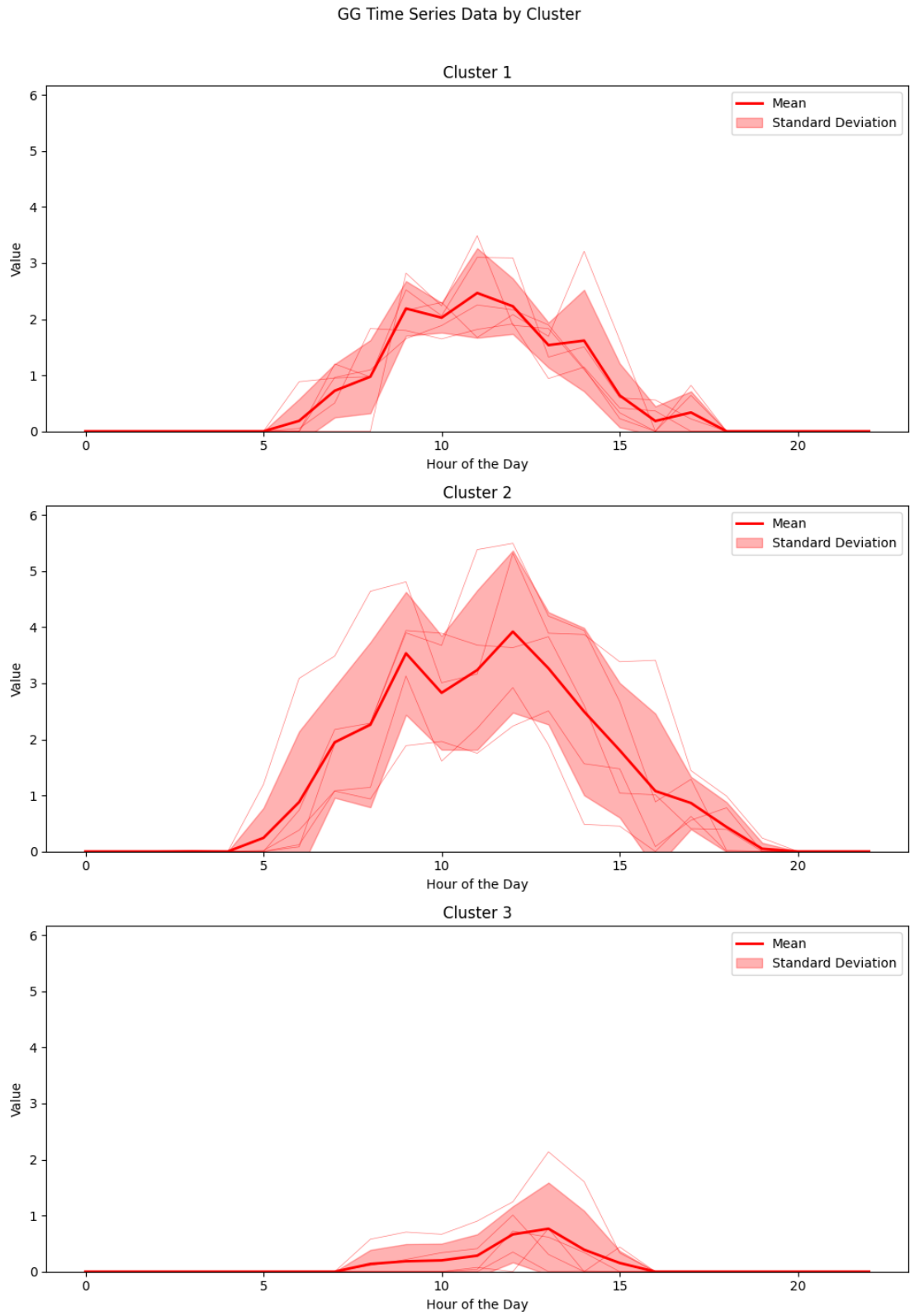


Figure 4.2: GG Time Series Data by Cluster

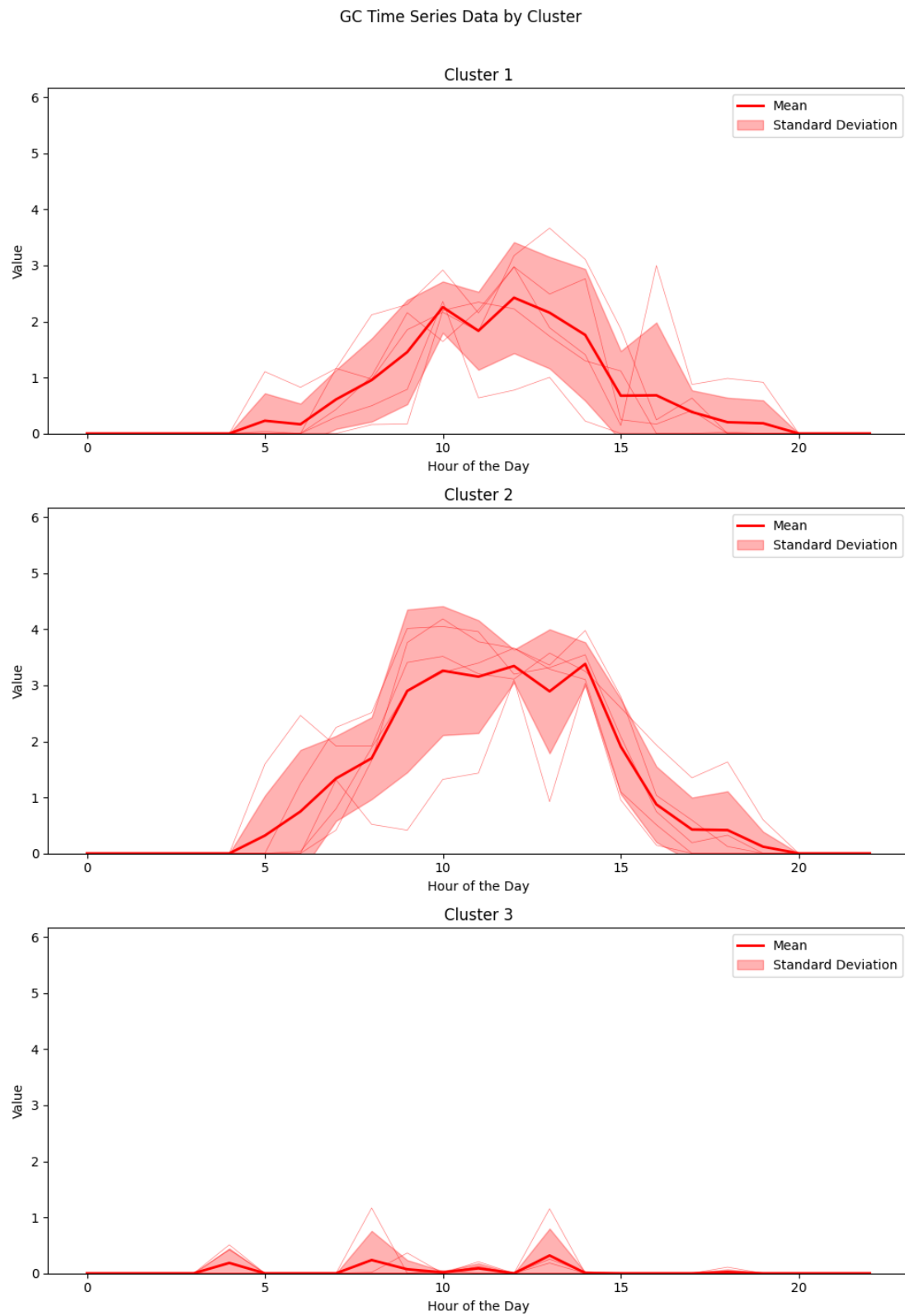


Figure 4.3: GC Time Series Data by Cluster

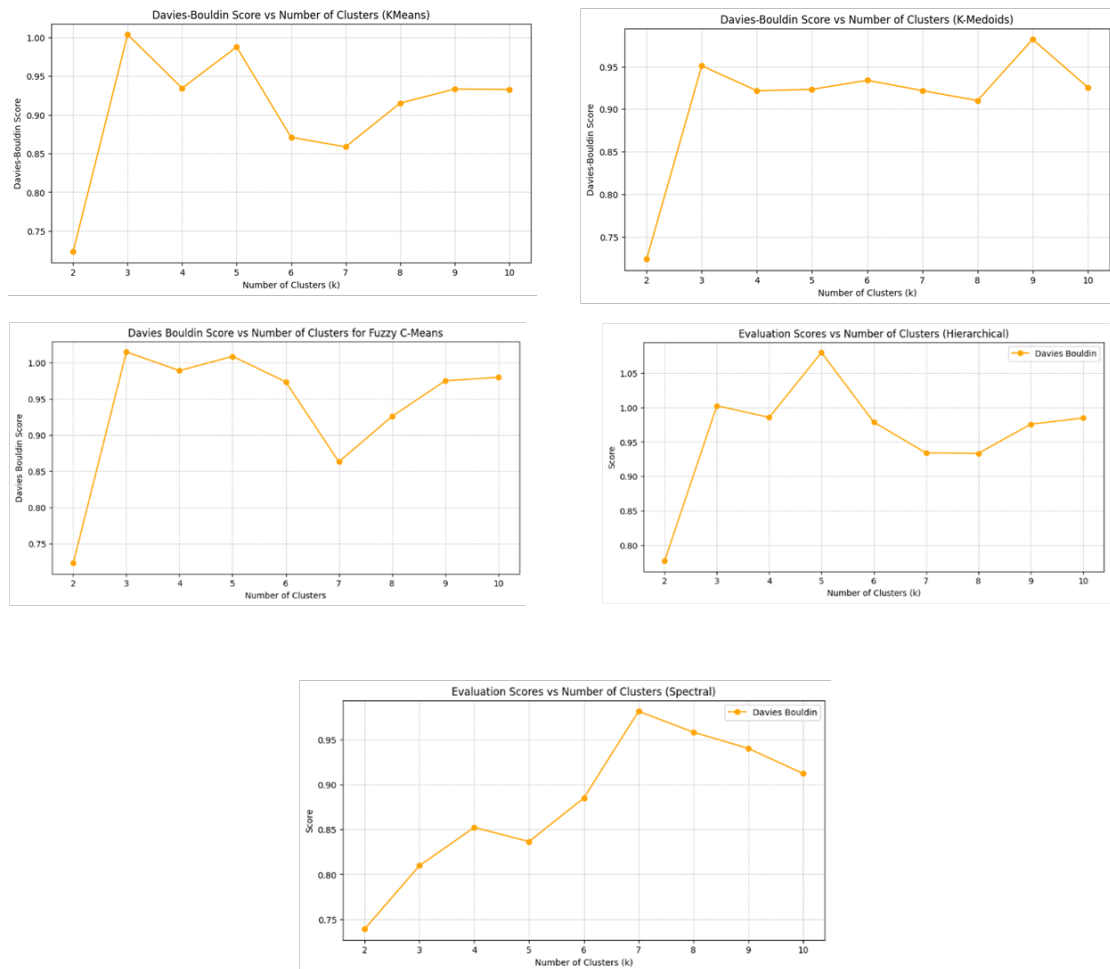


Figure 4.4: Davies-Bouldin Score vs Number of Clusters for Different Algorithms

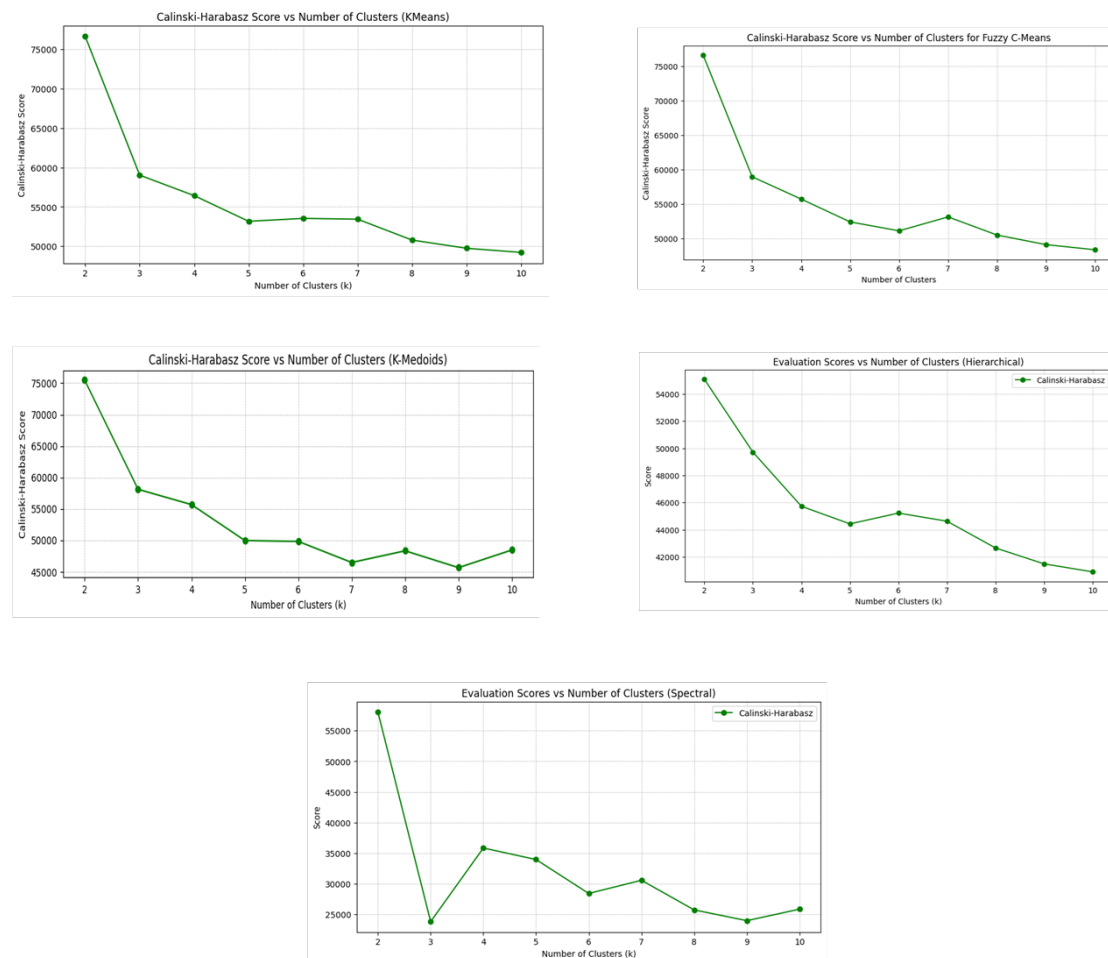


Figure 4.5: Calinski-Harabasz Score vs Number of Clusters for Different Algorithms

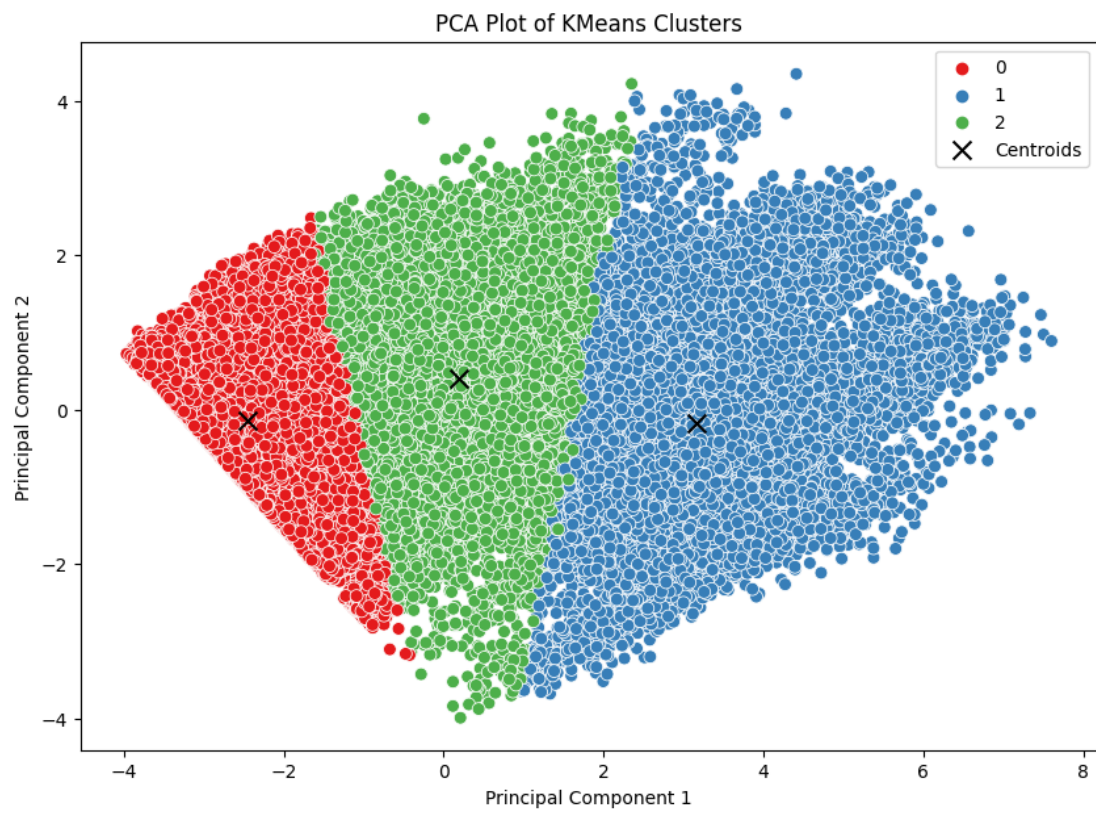


Figure 4.6: PCA Plot of KMeans Clusters

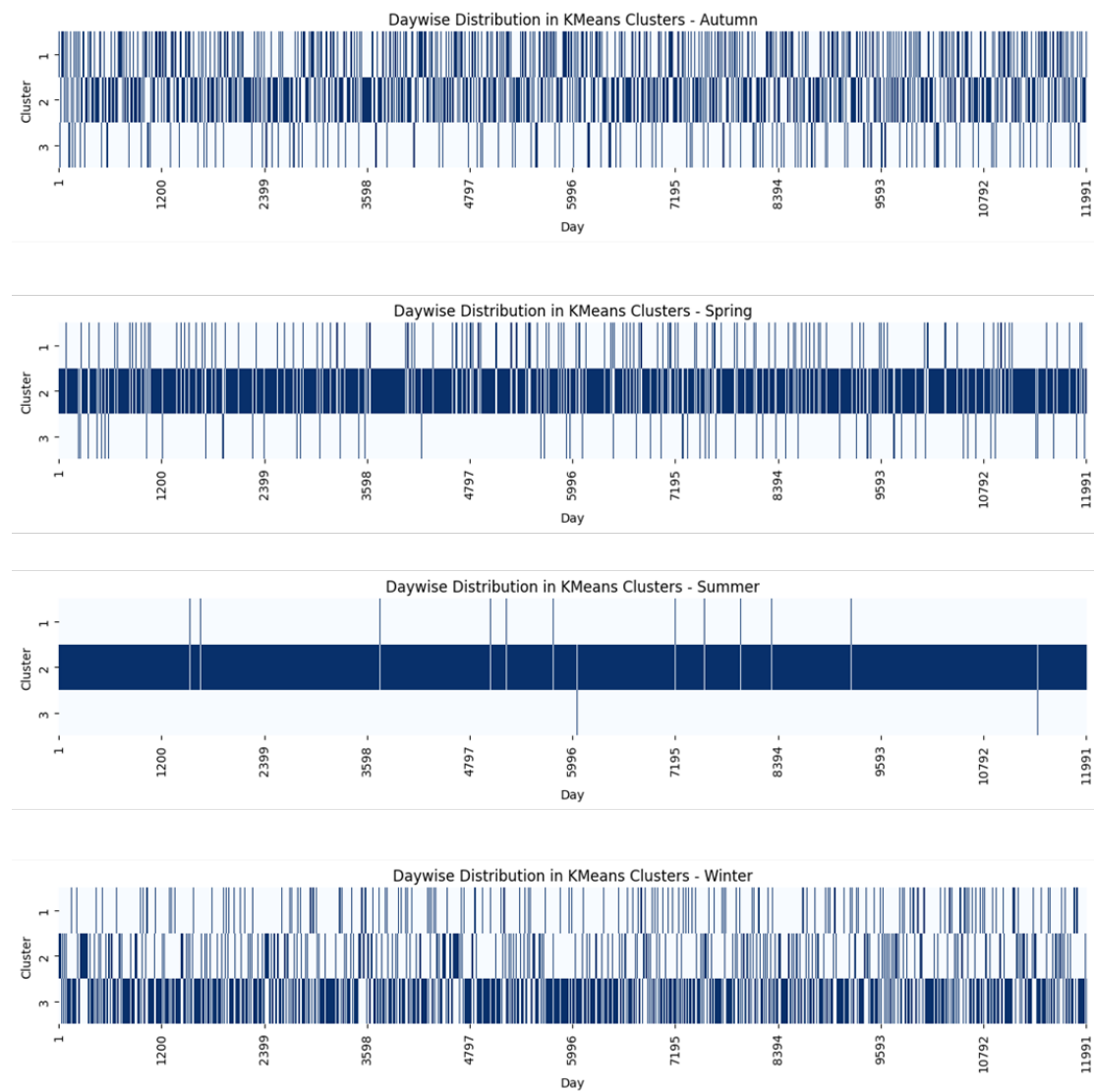


Figure 4.7: Daywise Distribution of KMeans Clusters across Seasons

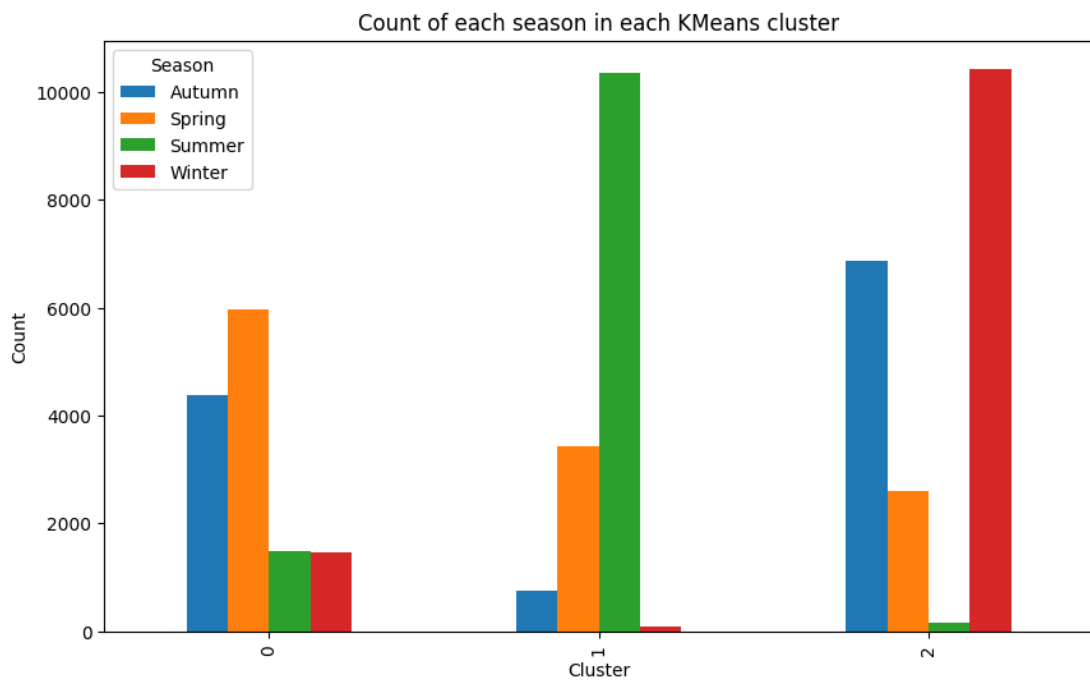


Figure 4.8: Count of Each Season in Each KMeans Cluster

Future Work and Conclusion

While this study provides significant insights into prosumer clustering, there are several areas for future research to enhance the findings and applications.

Scope for Future Works

1. **Exploration of Additional Clustering Algorithms:** While K-Means, Fuzzy C-Means, Hierarchical Clustering, K-Medoids, and Spectral Clustering were evaluated in this study, future work could explore other clustering algorithms such as DBSCAN and Gaussian Mixture Models. These algorithms may offer improved accuracy or interpretability over the ones used in this study, particularly for detecting clusters with irregular shapes or densities.
2. **Advanced Feature Extraction Techniques:** This study primarily focused on time window-based feature extraction and basic statistical measures for clustering. Future research could investigate advanced feature extraction methods such as wavelet transforms, deep learning-based feature extraction, and other sophisticated statistical techniques. These methods might capture more intricate patterns in energy consumption and generation behaviors, potentially leading to better clustering performance.
3. **Real-Time Integration with Energy Management Systems:** The integration of clustering results into real-time energy management systems could allow for dynamic adjustments to demand response (DR) programs based on changing prosumer behaviors. This could enhance the efficiency

and responsiveness of energy management, especially during peak load periods. Implementing real-time data streams and feedback loops would be a critical aspect of this integration.

4. **Longitudinal Analysis:** Conducting longitudinal studies to analyze how prosumer behaviors and clustering outcomes change over extended periods could provide deeper insights into the stability and evolution of energy consumption patterns. This study used datasets spanning multiple years, but future research could extend this to even longer durations and include more frequent data updates to track temporal changes more effectively.
5. **Impact of External Factors:** Future research could delve deeper into the impact of external factors such as economic conditions, policy changes, technological advancements, and detailed weather conditions on prosumer behaviors and clustering results. While this study considered seasonal effects, incorporating more granular data on external influences could enhance the understanding of factors driving energy usage patterns.
6. **Scalability and Generalization:** Testing the scalability of the proposed methodologies on larger and more diverse datasets, and validating their generalization across different geographic regions or types of energy systems, would be essential. This study utilized datasets from Australia (Ausgrid) and Denmark, but extending this to a global context would help in assessing the robustness and applicability of the findings.
7. **User-Centric DR Programs:** Developing and testing user-centric DR programs based on the clustering results could enhance program effectiveness and user satisfaction. The distinct clustering patterns identified in this study could inform the design of more personalized and efficient DR strategies, tailored to the specific needs and behaviors of each cluster.
8. **Incorporation of Weather Data:** Future studies could incorporate detailed weather data, such as temperature, solar irradiance, and humidity, to better understand their impact on energy consumption and generation. While seasonal variations were considered, more granular weather data could provide additional insights and improve clustering accuracy.
9. **Advanced Visualization Techniques:** Employing advanced visualization techniques could help in better understanding the clustering results and making the data more interpretable for stakeholders. This study utilized PCA plots, time series visualizations, and bar charts, but future work could

explore interactive dashboards, 3D visualizations, and other innovative tools to present data more effectively.

10. **Exploration of Consumer Behavior:** Integrating data on consumer behavior, such as appliance usage patterns, lifestyle information, and socio-demographic factors, could enhance the understanding of underlying drivers of energy usage. This could improve the accuracy of clustering and provide more actionable insights for DR program design.
11. **Enhanced Data Quality and Processing Techniques:** Future work could focus on improving data quality and developing more sophisticated data processing techniques to handle missing or noisy data. Ensuring high-quality data is crucial for accurate clustering and reliable results.
12. **Economic and Environmental Impact Analysis:** Analyzing the economic and environmental impacts of different clusters and DR programs could provide valuable insights for policymakers and energy providers. This could involve cost-benefit analyses, carbon footprint assessments, and evaluations of the long-term sustainability of various energy management strategies.

In conclusion, while this study lays a strong foundation for understanding prosumer energy behaviors through clustering, there are numerous opportunities for further research and practical applications. By continuing to explore and refine these techniques, we can advance the field of smart grid and energy management systems, ultimately contributing to a more efficient and sustainable energy future.

Bibliography

- [1] M. Ahmed, R. Seraj, and S. M. S. Islam, “The k-means algorithm: A comprehensive survey and performance evaluation,” *Electronics*, vol. 9, no. 8, p. 1295, 2020.
- [2] K. P. Sinaga and M.-S. Yang, “Unsupervised k-means clustering algorithm,” *IEEE access*, vol. 8, pp. 80 716–80 727, 2020.
- [3] S. Na, L. Xumin, and G. Yong, “Research on k-means clustering algorithm: An improved k-means clustering algorithm,” in *2010 Third International Symposium on intelligent information technology and security informatics*, Ieee, 2010, pp. 63–67.
- [4] J. Nayak, B. Naik, and H. Behera, “Fuzzy c-means (fcm) clustering algorithm: A decade review from 2000 to 2014,” in *Computational Intelligence in Data Mining-Volume 2: Proceedings of the International Conference on CIDM, 20-21 December 2014*, Springer, 2015, pp. 133–149.
- [5] J. C. Bezdek, R. Ehrlich, and W. Full, “Fcm: The fuzzy c-means clustering algorithm,” *Computers & geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.
- [6] H. Sun, S. Wang, and Q. Jiang, “Fcm-based model selection algorithms for determining the number of clusters,” *Pattern recognition*, vol. 37, no. 10, pp. 2027–2037, 2004.
- [7] N. K. Kaur, U. Kaur, and D. Singh, “K-medoid clustering algorithm-a review,” *Int. J. Comput. Appl. Technol*, vol. 1, no. 1, pp. 42–45, 2014.
- [8] S. Harikumar and P. Surya, “K-medoid clustering for heterogeneous datasets,” *Procedia Computer Science*, vol. 70, pp. 226–237, 2015.
- [9] H.-S. Park and C.-H. Jun, “A simple and fast algorithm for k-medoids clustering,” *Expert systems with applications*, vol. 36, no. 2, pp. 3336–3341, 2009.
- [10] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: An overview,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

- [11] F. Nielsen and F. Nielsen, “Hierarchical clustering,” *Introduction to HPC with MPI for Data Science*, pp. 195–211, 2016.
- [12] C. F. Olson, “Parallel algorithms for hierarchical clustering,” *Parallel computing*, vol. 21, no. 8, pp. 1313–1325, 1995.
- [13] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: An overview, ii,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 6, e1219, 2017.
- [14] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and computing*, vol. 17, pp. 395–416, 2007.
- [15] H. Jia, S. Ding, X. Xu, and R. Nie, “The latest research progress on spectral clustering,” *Neural Computing and Applications*, vol. 24, pp. 1477–1486, 2014.
- [16] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” *Advances in neural information processing systems*, vol. 14, 2001.
- [17] F. Bach and M. Jordan, “Learning spectral clustering,” *Advances in neural information processing systems*, vol. 16, 2003.
- [18] G. Ivosev, L. Burton, and R. Bonner, “Dimensionality reduction and visualization in principal component analysis,” *Analytical chemistry*, vol. 80, no. 13, pp. 4933–4944, 2008.
- [19] M. Fontes and C. Soneson, “The projection score—an evaluation criterion for variable subset selection in pca visualization,” *BMC bioinformatics*, vol. 12, pp. 1–17, 2011.
- [20] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [21] E. Becht, L. McInnes, J. Healy, *et al.*, “Dimensionality reduction for visualizing single-cell data using umap,” *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.
- [22] C. Shao and H. Huang, “Improvement of data visualization based on isomap,” in *Mexican International Conference on Artificial Intelligence*, Springer, 2005, pp. 534–543.
- [23] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, “Non-linear dimensionality reduction techniques for classification and visualization,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 645–651.

-
- [24] K. R. Shahapure and C. Nicholas, “Cluster quality analysis using silhouette score,” in *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, IEEE, 2020, pp. 747–748.
 - [25] J. Baarsch and M. E. Celebi, “Investigation of internal validity measures for k-means clustering,” in *Proceedings of the international multiconference of engineers and computer scientists*, International Association of Engineers Hong Kong, vol. 1, 2012, pp. 14–16.
 - [26] F. Ros, R. Riad, and S. Guillaume, “Pdbi: A partitioning davies-bouldin index for clustering evaluation,” *Neurocomputing*, vol. 528, pp. 178–199, 2023.
 - [27] T. F. Gonzalez, “Clustering to minimize the maximum intercluster distance,” *Theoretical computer science*, vol. 38, pp. 293–306, 1985.
 - [28] L. Sahu and B. R. Mohan, “An improved k-means algorithm using modified cosine distance measure for document clustering using mahout with hadoop,” in *2014 9th International Conference on Industrial and Information Systems (ICIIS)*, IEEE, 2014, pp. 1–5.
 - [29] K. Y. Yeung and W. L. Ruzzo, “Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data,” *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.
 - [30] S. Romano, J. Bailey, V. Nguyen, and K. Verspoor, “Standardized mutual information for clustering comparisons: One step further in adjustment for chance,” in *International conference on machine learning*, PMLR, 2014, pp. 1143–1151.
 - [31] A. K. Kar, A. C. Mishra, and S. K. Mohanty, “An efficient entropy based dissimilarity measure to cluster categorical data,” *Engineering Applications of Artificial Intelligence*, vol. 119, p. 105 795, 2023.
 - [32] C. Silva, P. Faria, Z. Vale, and N. Starzacher, “Aggregation of consumers and producers in a community with different clustering methods,” in *Fourth DREAM-GO Workshop-Demand response approaches for real-time renewable energy integration*, Instituto Politécnico do Porto. Instituto Superior de Engenharia, 2019.
 - [33] D. J. Vergados, I. Mamounakis, P. Makris, and E. Varvarigos, “Prosumer clustering into virtual microgrids for cost reduction in renewable energy trading markets,” *Sustainable Energy, Grids and Networks*, vol. 7, pp. 90–103, 2016.

- [34] L. Han, T. Morstyn, C. Crozier, and M. McCulloch, "Improving the scalability of a prosumer cooperative game with k-means clustering," in *2019 IEEE Milan PowerTech*, IEEE, 2019, pp. 1–6.
- [35] A. D. Rathnayaka, V. M. Potdar, and S. J. Kuruppu, "An innovative approach to manage prosumers in smart grid," in *2011 World Congress on Sustainable Technologies (WCST)*, IEEE, 2011, pp. 141–146.
- [36] S. Pelekis, A. Pipergias, E. Karakolis, *et al.*, "Targeted demand response for flexible energy communities using clustering techniques," *Sustainable Energy, Grids and Networks*, vol. 36, p. 101 134, 2023.
- [37] S. Mythili and E. Madhiya, "An analysis on clustering algorithms in data mining," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 1, pp. 334–340, 2014.
- [38] K. Bindra and A. Mishra, "A detailed study of clustering algorithms," in *2017 6th international conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO)*, IEEE, 2017, pp. 371–376.
- [39] M. Z. Rodriguez, C. H. Comin, D. Casanova, *et al.*, "Clustering algorithms: A comparative approach," *PloS one*, vol. 14, no. 1, e0210236, 2019.
- [40] E. A. Gupta, E. A. Gupta, and A. Mishra, "Research paper on cluster techniques of data variations," *International Journal of Advance Technology & Engineering Research*, vol. 1, no. 1, pp. 39–47, 2011.
- [41] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of data science*, vol. 2, pp. 165–193, 2015.
- [42] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104 743, 2022.
- [43] K. Mishra, S. Basu, and U. Maulik, "Load profile mining using directed weighted graphs with application towards demand response management," *Applied Energy*, vol. 311, p. 118 578, 2022.
- [44] O. Motlagh, A. Berry, and L. O'Neil, "Clustering of residential electricity customers using load time series," *Applied energy*, vol. 237, pp. 11–24, 2019.
- [45] L. G. B. Ruiz, M. Pegalajar, R. Arcucci, and M. Molina-Solana, "A time-series clustering methodology for knowledge extraction in energy consumption data," *Expert Systems with Applications*, vol. 160, p. 113 731, 2020.

- [46] A. Javed, B. S. Lee, and D. M. Rizzo, “A benchmark study on time series clustering,” *Machine Learning with Applications*, vol. 1, p. 100 001, 2020.
- [47] E. E. Özkoç, “Clustering of time-series data,” *Data Mining-Methods, Applications and Systems*, p. 87, 2020.
- [48] A. Alqahtani, M. Ali, X. Xie, and M. W. Jones, “Deep time-series clustering: A review,” *Electronics*, vol. 10, no. 23, p. 3001, 2021.
- [49] X. Wang, K. A. Smith, R. Hyndman, and D. Alahakoon, “A scalable method for time series clustering,” *Proc. Unrefereed Res. Papers*, vol. 1, 2004.
- [50] P. Skaloumpakas, E. Sarmas, M. Rachmanidis, and V. Marinakis, “Reshaping the energy landscape of crete through renewable energy valleys,” *Scientific Reports*, vol. 14, no. 1, p. 8038, 2024.
- [51] Y. Wang, Q. Chen, T. Hong, and C. Kang, “Review of smart meter data analytics: Applications, methodologies, and challenges,” *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2018.
- [52] U. H. Ramadhani, M. Shepero, J. Munkhammar, J. Widen, and N. Etherden, “Review of probabilistic load flow approaches for power distribution systems with photovoltaic generation and electric vehicle charging,” *International Journal of Electrical Power & Energy Systems*, vol. 120, p. 106 003, 2020.
- [53] D. Gebbran, S. Mhanna, Y. Ma, A. C. Chapman, and G. Verbič, “Fair coordination of distributed energy resources with volt-var control and pv curtailment,” *Applied Energy*, vol. 286, p. 116 546, 2021.
- [54] N. Liu, M. Cheng, X. Yu, J. Zhong, and J. Lei, “Energy-sharing provider for pv prosumer clusters: A hybrid approach using stochastic programming and stackelberg game,” *IEEE Transactions on Industrial Electronics*, vol. 65, no. 8, pp. 6740–6750, 2018.
- [55] D. Holtschulte, A. Erlangga, E. Ortjohann, *et al.*, “Local energy markets in clustering power system approach for smart prosumers,” in *2017 6th International Conference on Clean Electrical Power (ICCEP)*, IEEE, 2017, pp. 215–222.
- [56] D. L. Donaldson and D. Jayaweera, “Effective solar prosumer identification using net smart meter data,” *International Journal of Electrical Power & Energy Systems*, vol. 118, p. 105 823, 2020.
- [57] E. L. Ratnam, S. R. Weller, C. M. Kellett, and A. T. Murray, “Residential load and rooftop pv generation: An australian distribution network dataset,” *International Journal of Sustainable Energy*, vol. 36, no. 8, pp. 787–806, 2017.

- [58] R. Yuan, S. A. Pourmousavi, W. L. Soong, A. J. Black, J. A. Liisberg, and J. Lemos-Vinasco, “A synthetic dataset of danish residential electricity prosumers,” *Scientific Data*, vol. 10, no. 1, p. 371, 2023.