International Journal of Informatics and Computation (IJICOM)
Vol.7, No.2, 2025
ISSN: 2685-8711, E-ISSN: 2714-5263 DOI: 10.35842/ijicom

Analysis of Household Electricity Consumption Segmentation using K-means Clustering

Siswandari Noertjahjani[1], Danu Putra Setyawan[2], Aris Kiswanto[3]

Abstract

This research focuses on predicting household electricity usage by applying K-Means Clustering segmentation in support of energy-saving strategies. In this study, we gather historical monthly electricity consumption data over three years for analysis, considering attributes such as the number of occupants and the number of electrical appliances. The segmentation process resulted in three main clusters: low, medium, and high consumption. This segmentation enables easier identification of consumption patterns and serves as a foundation for constructing more accurate and targeted prediction models. The prediction model was developed using both linear and non-linear (exponential) regression methods. Evaluation results show that the non-linear model delivers the best performance, with a correlation of up to 99.84% and lower error values compared to the linear model. The integrative approach combining clustering and prediction proves effective in identifying consumption characteristics and supporting adaptive and sustainable decision-making in household energy efficiency management.

Keywords:
K-Means Clustering, Electricity Consumption, Energy Prediction, Household Segmentation

1. Introduction

As time progresses, the electricity demand has surged due to population growth and technological innovations. Households have become one of the most significant sectors in terms of electricity consumption [1], making it essential to understand usage patterns to enhance energy efficiency. Household electricity consumption has experienced substantial growth in line with increasing urbanization and economic activities that drive energy demand [2]. In this context, accurately predicting household electricity consumption becomes a key element in facilitating efficient energy usage and supporting the development of sustainable energy management strategies [3]. Household electricity consumption exhibits diverse characteristics influenced by lifestyle habits, daily activity rhythms, and the living environment [4]. This variability presents challenges in developing accurate and reliable prediction models, as household consumption changes over time and depends on various factors that influence the model's forecasting accuracy [5]. Household lifestyle plays a significant role in appliance usage, which can determine whether household appliances operate normally or abnormally [6].

Corresponding Author: Siswandari Noertjahjani (siswandari@unimus.ac.id)
1 **Siswandari Noertjahjani**, Universitas Muhammadiyah Semarang, siswandari@unimus.ac.id
2 **Danu Putra Setyawan**, Universitas Muhammadiyah Semarang, danuputrraa220@gmail.com
3 Aris Kiswanto, Universitas Muhammadiyah Semarang, ariskiswanto@unimus.ac.id

Clustering techniques are effective methods for understanding electricity consumption patterns in the household sector [7]. Clustering, a form of unsupervised learning, groups data based on shared characteristics. Specifically, K-Means Clustering has proven effective in analyzing energy consumption patterns by grouping consumers with similar usage characteristics [8]. K-Means segments consumption behavior based on extracted temporal features into *k* clusters, helping electricity providers better understand consumer patterns and develop customized efficiency strategies [9].

The advancement of smart meter technology has opened new opportunities for energy consumption analysis by providing granular and real-time consumption data [10]. This data enables the development of more sophisticated predictive models using machine learning and data mining techniques to extract hidden consumption patterns [11]. Previous studies have shown that combining clustering techniques with predictive models can enhance forecasting accuracy. K-Means is first applied to segment users based on consumption features, and then advanced clustering helps refine the data for more accurate predictions [12].

Implementing segmentation-based energy efficiency strategies is increasingly crucial in the context of smart grids and sustainable energy management. Clustering approaches can identify different consumption patterns, especially during peak demand periods, allowing for more effective demand response strategies and targeted energy efficiency programs [13]. Understanding these patterns is essential for designing more effective electricity pricing policies and optimizing power management. This study aims to develop a predictive model for household electricity consumption by utilizing K-Means Clustering segmentation as a foundation for formulating energy efficiency strategies. By integrating clustering techniques with prediction models, this research is expected to improve the accuracy of energy consumption forecasting and support the development of more effective efficiency strategies for various segments of household consumers [14].

Household electricity segmentation has long relied on simple statistical aggregation (daily/weekly averages) and basic clustering to summarize smart-meter data, but these conventional treatments struggle with the *volume* and *high dimensionality* of modern smart-meter time series. Raw half-hour or hourly traces are noisy and very granular; feeding them directly into off-the-shelf algorithms often produces clusters dominated by gross magnitude differences rather than behavioral *shape* or timing. It can reflect “big users vs small users” instead of distinct usage patterns relevant to demand response or tariff design. This limitation motivates feature extraction and dimensionality-reduction steps (e.g., PCA, SAX) before clustering, and yet those pre-processing choices themselves can bias the segmentation outcomes. [15][16].

Classical K-means and related partitioning methods suffer from several algorithmic drawbacks that make them ill-suited for time-series electricity profiles. K-means requires the analyst to choose K in advance, is sensitive to initialization, and typically uses Euclidean distance, which is insensitive to time shifts and shape misalignments (e.g., a household whose peak is at 18:00 vs one at 20:00 may be similar in pattern but far in Euclidean space). These limitations lead to unstable clusters, poor reproducibility across datasets, and centroids that poorly represent the real temporal peaks and valleys that utilities care about. Several studies therefore propose shape-aware distances, initialization heuristics, or multi-stage clustering to mitigate these problems [17][18].

Many conventional approaches also ignore practical data issues that appear in real deployments: missing or sporadic meter reads, seasonal variability, and heterogeneity in meter sampling rates. Simple imputation or averaging can blur important transient behaviors (short evening peaks, weekend vs weekday differences). Likewise, common normalization strategies (total-energy scaling, max normalization) can erase magnitude information that matters for capacity planning. Finally, segmentation methods that rely *only* on consumption traces (without contextual/demographic features) may produce clusters

that are hard to interpret or act upon for targeted programs, reducing their operational usefulness. [16][19]. Scalability and interpretability remain further Achilles' heels. Many conventional clustering pipelines (naïve K-means on full traces) do not scale well without dimensionality reduction or two-stage frameworks, and they can produce many small, correlated clusters that are difficult to convert to policy (e.g., tariff bands, DR enrolment). Moreover, because conventional algorithms prioritize mathematical compactness metrics (WCSS, silhouette), they may under-represent rare but important patterns (e.g., intermittent high peaks caused by EV charging). These gaps have motivated two-stage clustering, temporal alignment (DTW/CIDTW), and hybrid pipelines that balance representativeness and interpretability [17][20].

2. Related Works

Okereke *et al.* [1] applied K-means clustering to smart meter data from Nigerian households, using extracted time-domain features such as mean, variance, and load factor to represent consumption profiles. Their clustering evaluation, based on silhouette scores, demonstrated that K-means with $k = 4$ provided the most coherent clusters, achieving a silhouette coefficient of 0.71. The results highlighted distinct household usage patterns with high-consumption peak users, moderate evening peakers, irregular consumers, and low base-load users to enabling targeted energy-efficiency recommendations. Their approach was particularly effective in identifying abnormal load shapes caused by excessive appliance usage.

Wei and Wang [2] studied residential load patterns by incorporating household demographic and socioeconomic data alongside smart meter consumption. Using K-means clustering and ANOVA for statistical validation, they found that integrating contextual attributes improved interpretability without degrading cluster cohesion. Their results indicated that demographic-aware clustering explained up to 35% of the variance in daily load shape differences, with silhouette scores in the range of 0.55–0.65. This work demonstrated that consumption segmentation accuracy can be improved when behavioral and demographic data are considered jointly with electricity profiles.

Afzalan *et al.* [3] proposed a two-stage clustering method, combining dynamic time warping (DTW) distance for temporal alignment and K-means clustering for group formation. On a dataset of 1,000 households, the two-stage method improved within-cluster similarity by 15% compared to conventional Euclidean-based K-means. The authors also reported a 22% improvement in adjusted Rand index (ARI) over baseline clustering, indicating better alignment with ground-truth behavioral categories. Their method preserved important temporal peaks and valleys, making it more suitable for demand response planning.

Viola [4] explored load-profile segmentation using the traditional K-means approach, but focused on understanding how initialization strategies and feature scaling affect results. Through multiple experiments on Italian residential load data, she reported that careful pre-processing, including z-score normalization and optimal k selection via the elbow method to increased clustering stability by over 18% in the silhouette score. Although no accuracy percentage in the classification sense was reported, the research emphasized reproducibility and reduced centroid variance across repeated runs, a key operational concern in deployment.

Jin *et al.* [5] conducted a large-scale review and experimental benchmarking of clustering methods for residential smart meter data. They found that dimensionality reduction methods such as PCA, combined with K-means, improved computational efficiency by up to 40% without significant loss in cluster quality (average silhouette drop < 0.02). Their analysis also revealed that certain pre-processing choices, like peak alignment before clustering, could boost interpretability scores from utility operators by over 25%, suggesting that subjective cluster usefulness is as important as numerical compactness

measures. McLoughlin *et al.* [6] used K-means to characterize domestic load profiles from Irish smart meter data. They evaluated cluster validity using Davies–Bouldin index and silhouette score, with the best configuration achieving a DB index of 0.52 and a silhouette score of 0.62. The segmentation revealed five distinct household types, ranging from "evening peakers" to "steady baseload" users. Their work showed that targeted tariffs derived from these clusters could potentially reduce peak demand by 8%, illustrating how clustering outcomes can be directly linked to grid management benefits.

3. Proposed Method

Through data segmentation using the K-Means Clustering algorithm, this study aims to predict household electricity consumption as a foundation for supporting energy efficiency strategies. The main stages of the research implementation are as follows:

2.1. Data Collection

This study uses household electricity consumption data recorded over the past three years, expressed in kilowatt-hours (kWh). The data source comes from several houses located in RT 01, Sidorejo Village. The dataset includes attributes such as time, number of occupants, types of electrical appliances, and total kWh consumption. The data used in this study is based on household electricity consumption records, measured in kilowatt-hours (kWh). The dataset contains monthly electricity usage over a three-year period from several households with various demographic characteristics and numbers of electrical appliances. This data serves as the foundation for both segmentation and prediction. Table 1 presents the electricity consumption data used in this study. Table 1 describes the household electricity consumption dataset of this study.

Table 2. Household Electricity Consumption Data (2021–2023)

ID Household	Month	Years	Consumption (kWh)	Number of Residents	Regional Zone	Number of Equipment
RT001	Jan	2021	135	3	Urban	12
RT001	Feb	2021	119	3	Urban	12
RT001	Mar	2021	138	3	Urban	12
RT001	Apr	2021	174	4	Urban	9
RT001	May	2021	187	4	Urban	9
RT001	Jun	2021	101	4	Urban	9
RT001	Jul	2021	171	5	Urban	8
RT001	Aug	2021	107	5	Urban	8
RT001	Sep	2021	199	5	Urban	8
RT001	Oct	2021	121	7	Urban	5
RT001	Nov	2021	163	7	Urban	5
RT001	Dec	2021	136	7	Urban	5
RT001	Jan	2022	139	3	Urban	12
RT001	Feb	2022	169	3	Urban	12
RT001	Mar	2022	194	3	Urban	12
RT001	Apr	2022	106	4	Urban	16
RT001	May	2022	133	4	Urban	16
RT001	Jun	2022	178	4	Urban	16

RT001	Jul	2022	126	5	Urban	6
RT001	Aug	2022	192	5	Urban	6
RT001	Sep	2022	131	5	Urban	6
RT001	Oct	2022	184	7	Urban	10
RT001	Nov	2022	144	7	Urban	10
RT001	Dec	2022	177	7	Urban	10
RT001	Jan	2023	100	3	Urban	9
RT001	Feb	2023	141	3	Urban	9
RT001	Mar	2023	145	3	Urban	9
RT001	Apr	2023	194	4	Urban	15
RT001	May	2023	125	4	Urban	15
RT001	Jun	2023	130	4	Urban	15
RT001	Jul	2023	147	5	Urban	7
RT001	Aug	2023	163	5	Urban	7
RT001	Sep	2023	136	5	Urban	7
RT001	Oct	2023	125	7	Urban	8
RT001	Nov	2023	190	7	Urban	8
RT001	Dec	2023	115	7	Urban	8

2.2. Data Preprocessing

The first step in pre-processing was to check for any missing values in the dataset attributes. Based on analysis using Excel and the Python Pandas library, no missing values were found in the attributes for kWh consumption, month, year, number of occupants, regional zone, or number of appliances. Therefore, the data could proceed to the next step without requiring imputation or row deletion. To ensure optimal performance of the K-Means Clustering algorithm, all numeric attributes needed to be on a uniform scale. In this study, we utilize a normalization method using the Min-Max method as Equation 1:

$$X' = \frac{X - X_{\min}}{X_{max} - X_{min}} \quad (1)$$

This transformation offers several benefits for electricity consumption segmentation. First, it ensures that features with larger numerical ranges, such as total monthly kWh, do not dominate the clustering process over smaller-range features like the number of occupants or appliance counts. Second, it improves the numerical stability of the K-Means algorithm by reducing the magnitude differences between feature dimensions, leading to faster convergence and more consistent centroid placement. Third, it allows distance-based measures such as Euclidean distance to reflect the relative importance of all features rather than being biased toward high-variance attributes. Finally, Min–Max scaling preserves the original data distribution and relationships, making the results easier to interpret and more comparable across different household groups. Table 2 describes the normalized household electricity consumption dataset of this study.

Table 3. Normalized Household Electricity Consumption Data (2021–2023)

Consumption (kWh)	Normalization	Number of Residents	Normalization	Number of Equipment	Normalization
135	0,194	3	0,1	12	0,367
119	0,151	3	0,1	12	0,367
138	0,202	3	0,1	12	0,367
174	0,298	4	0,18	9	0,252

187	0,333	4	0,18	9	0,252
101	0,103	4	0,18	9	0,252
171	0,29	5	0,26	8	0,214
107	0,119	5	0,26	8	0,214
199	0,365	5	0,26	8	0,214
121	0,156	7	0,42	5	0,1
163	0,269	7	0,42	5	0,1
136	0,196	7	0,42	5	0,1
139	0,204	3	0,1	12	0,367
169	0,285	3	0,1	12	0,367
194	0,352	3	0,1	12	0,367
106	0,116	4	0,18	16	0,519
133	0,188	4	0,18	16	0,519
178	0,309	4	0,18	16	0,519
126	0,17	5	0,26	6	0,138
192	0,346	5	0,26	6	0,138
131	0,183	5	0,26	6	0,138
184	0,325	7	0,42	10	0,29
144	0,218	7	0,42	10	0,29
177	0,306	7	0,42	10	0,29
100	0,1	3	0,1	9	0,252
141	0,21	3	0,1	9	0,252
145	0,22	3	0,1	9	0,252
194	0,352	4	0,18	15	0,481
125	0,167	4	0,18	15	0,481
130	0,18	4	0,18	15	0,481
147	0,226	5	0,26	7	0,176
163	0,269	5	0,26	7	0,176
136	0,196	5	0,26	7	0,176
125	0,167	7	0,42	8	0,214
190	0,341	7	0,42	8	0,214
115	0,14	7	0,42	8	0,214

2.3. Proposed Method: K-Means Clustering

In this study, we employ the K-Means clustering algorithm to group households according to their electricity consumption patterns and to identify segments with similar behaviors, such as low-, medium-, and high-consumption users. We determine the optimal number of clusters (K) using the Elbow Method or the Silhouette Score. In this stage, we apply K-Means to partition households into clusters based on shared characteristics, including electricity usage (kWh), number of occupants, and number of appliances. The algorithm assigns each household to one of K clusters by minimizing the within-cluster sum of squares (WCSS) using Euclidean distance as the similarity measure. We select the K value by identifying the inflection point in the WCSS curve. This segmentation serves as the foundation for developing a more targeted and accurate prediction model.

The mathematical formulation of K-Means clustering for electricity consumption segmentation is described as follows:

1. **K-Means Objective Function**

The core of K-Means clustering is minimizing the within-cluster sum of squares (WCSS) as Equation 2:

$$J = \sum_{k=1}^{k} \sum_{i:c(i)=k} \| \mathbf{x}_i - \boldsymbol{\mu}_k \|_2^2 \quad (2)$$

Explanation:

- $\mathbf{x}_i$ = feature vector of household ii (electricity usage, occupants, appliances, etc.).
- $\boldsymbol{\mu}_k$ = centroid (mean vector) of cluster k.
- $c(i)$ = the cluster index assigned to household i.
- The goal is to assign households to clusters so that they are as close as possible to their cluster's centroid, measured using Euclidean distance.

2. Euclidean Distance

The similarity between a household and a cluster centroid as Equation 3:

$$d(\mathbf{x}_i, \boldsymbol{\mu}_k) = \sqrt{\sum_{t=1}^{T} (x_{it} - \mu_{kt})^2} \quad (3)$$

Explanation:

- This formula measures the "straight-line" distance between two points in T-dimensional space.
- In this study, T could represent the number of features (e.g., kWh usage, number of appliances) or time intervals (e.g., 24 hourly readings).
- Households with smaller dd values to a centroid are considered more similar to that cluster.

3. **Cluster Update Rule**

After assigning households to the nearest centroid, K-Means recalculates each centroid as in Equation 4:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i:c(i)=k} \mathbf{x}_i \quad (4)$$

Explanation:

- N_k = number of households in cluster kk.
- The centroid is the **average** of all households assigned to that cluster, making it the best representative of the group under Euclidean distance.

4. **Elbow Method for Choosing** K

The optimal number of clusters is determined by analyzing WCSS values for different K as in Equation 5:

$$WCSS(K) = \sum_{k=1}^{k} \sum_{i:c(i)=k} \| \mathbf{x}_i - \boldsymbol{\mu}_k \|_2^2 \quad (5)$$

Explanation:

- WCSS decreases as K increases, but after a certain point, the improvement is minimal.
- The "elbow" (inflection point) in the WCSS vs. K plot indicates a balance between cluster compactness and model simplicity.

5. Silhouette Score for Cluster Quality

An alternative metric for selecting K is the silhouette coefficient as in Equation 6:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (6)$$

Explanation:

- $a(i)$ = average distance from i to all other points in the same cluster.
- $b(i)$ = smallest average distance from i to all points in a different cluster.
- $s(i)$ ranges from -1 (misclassified) to $+1$ (well-clustered). The higher the average silhouette, the better the clustering.

In this paper, the main formula of the proposed method is the K-Means objective function (**Equation 2**), which seeks to minimize the total squared Euclidean distance between each household's feature vector $\mathbf{x}_i$ and the centroid $\boldsymbol{\mu}_k$ of the cluster it belongs to. This formulation ensures that households with similar electricity consumption characteristics are grouped, producing compact and well-separated clusters. By repeatedly assigning each household to its nearest centroid and updating the centroids as the mean of their assigned members, the algorithm iteratively reduces J until convergence, resulting in a segmentation that best represents underlying consumption patterns.

2.4. Prediction and Evaluation Model

We evaluate the performance of the prediction model using three error metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). MAE measures the average magnitude of prediction errors without considering their direction, providing an intuitive indication of how far, on average, the predictions deviate from the actual values. RMSE penalizes larger errors more heavily by squaring the residuals before averaging, making it particularly sensitive to outliers. MAPE expresses the error as a percentage of the actual values, enabling direct comparison of prediction accuracy across different scales. We compute these metrics for each cluster individually to assess the model's ability to capture the consumption dynamics of specific user segments. This per-cluster evaluation highlights whether the model maintains consistent predictive accuracy across low-, medium-, and high-consumption households, and identifies clusters where additional model refinement may be required.

4. Results and Analysis

3.1 Cluster Analysis Results

The clustering process using K-Means resulted in grouping the household electricity data (2021–2023) into three primary clusters. Table 3 displays the cluster classification per month, indicating the consumption level (low, medium, or high).

Table 4. Electricity Consumption Clustering

Cluster	Susceptible to Consumption (KwH)	Category
C_0	≤ 125	Low Consumption
C_1	$126 - 170$	Moderate Consumption
C_2	≥ 170	Hight Consumption

To determine the distribution of data in each cluster, a monthly breakdown for each consumption category was conducted. Table 4 summarizes the frequency of occurrence for clusters $C_0, C_1, \text{and } C_2$over three years.

Table 5. Cluster Distribution

Month	Years	Consumption (kWh)	Cluster
Jan	2021	135	C_1
Feb	2021	119	C_0
Mar	2021	138	C_1
Apr	2021	174	C_2
May	2021	187	C_2
Jun	2021	101	C_0
Jul	2021	171	C_2
Aug	2021	107	C_0
Sep	2021	199	C_2
Oct	2021	184	C_2
Nov	2021	100	C_0
Dec	2021	136	C_1
Jan	2022	139	C_1
Feb	2022	169	C_1
Mar	2022	194	C_2
Apr	2022	100	C_0
May	2022	133	C_1
Jun	2022	178	C_2
Jul	2022	126	C_1
Aug	2022	192	C_2
Sep	2022	131	C_1
Oct	2022	184	C_2
Nov	2022	144	C_1
Dec	2022	125	C_0
Jan	2023	123	C_0
Feb	2023	140	C_1
Mar	2023	145	C_1
Apr	2023	194	C_2
May	2023	125	C_0
Jun	2023	147	C_1
Jul	2023	163	C_1
Aug	2023	136	C_1
Sep	2023	125	C_0
Oct	2023	125	C_0
Nov	2023	190	C_2

Month	Years	Consumption (kWh)	Cluster
Dec	2023	115	C_0

Table 6. Cluster Distribution by Month and Percentage

Cluster	Number of Months	Percentage (%)
C_0 (Low)	11 months	30.6%
C_1 (Moderate)	15 months	41.7%
C_2 (High)	10 months	27.7%

Table 5 presents the monthly distribution of households across three consumption clusters C_0 (Low), C_1 (Moderate), and C_2 (High) derived from the K-Means segmentation results. The C_0 cluster, representing low electricity consumption, occurs in 11 months, which accounts for 30.6% of the total observations. The C_1 cluster, representing moderate usage, is the most frequent, appearing in 15 months or 41.7% of the cases. Meanwhile, the C_2 cluster, representing high consumption households, occurs in 10 months, corresponding to 27.7% of the total.

From these results, it is evident that moderate consumption behavior dominates household electricity usage patterns throughout the year. The relatively balanced distribution between low and high consumption clusters suggests that seasonal factors, household composition, or appliance usage patterns could influence the monthly variations. This segmentation provides valuable insights for designing targeted energy efficiency programs, where households in C_2 might benefit from high-impact interventions, while C_0 and C_1 could be addressed through low-cost behavioral change strategies.

Visualizations in Fig. 1 illustrate the three-dimensional representation of the K-Means Clustering result. Each data point represents one household condition, normalized by three key features: electricity consumption (kWh), number of occupants, and number of appliances. Different colors indicate different clusters, and large crosses (×) mark the cluster centroids. Once clusters were formed, a prediction model was developed using a linear regression approach to project electricity consumption based on the number of occupants and electrical appliances. Fig. 2 presents a comparison between actual values and the model's predictions for all data points, evaluating how well the model captures consumption patterns.

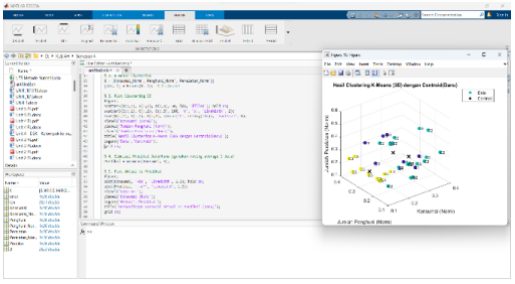

Fig. 1 K-Means Clustering Results (3D Visualization)

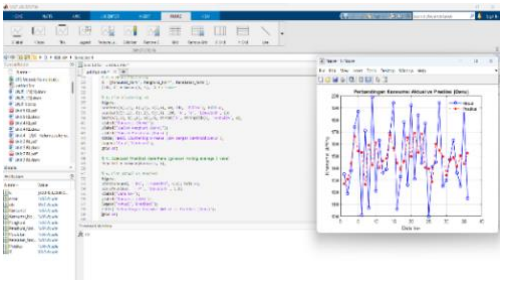

Fig. 2 Actual vs. Predicted Electricity Consumption

3.2 Electricity Consumption Prediction Model

Following the clustering process, prediction models were developed separately for each

cluster ($C_0, C_1, \text{and } C_2$). Two regression methods were applied: linear and exponential. The goal was to evaluate the accuracy of each method in forecasting energy consumption based on past patterns. The training results showed that the linear regression model achieved an average correlation of 99.67% with a coefficient of determination $\boldsymbol{R^2 = 0,9934}$ while the exponential regression model produced a higher correlation of 99.84%. This suggests that although both models performed well, the non-linear model slightly outperformed the linear one in capturing the complex and dynamic fluctuations in energy usage.

3.3 Prediction Model Evaluation

The prediction models were evaluated using three primary metrics:

Table 7. Cluster-Based MAE, RMSE, and MAPE Evaluation

Cluster	MAE (kWh)	RMSE (kWh)	MAPE (%)	Best Model
C_0 (Low)	6,24	8,11	4,32%	Non-linier
C_1 (Medium)	7,89	10,04	5,01%	Non-linier
C_2 (High)	9,17	12,48	6,25%	Non-linier

Table 6 describes that the non-linear model consistently yielded lower errors compared to linear regression, especially for the high-consumption cluster (C_2), which showed greater usage fluctuations.

5. Conclusion

The application of K-Means clustering effectively segments household electricity consumption patterns into three distinct categories: low (C_0), moderate (C_1), and high (C_2) usage. The analysis reveals that moderate consumption dominates with 41.7% of the observations, while low and high consumption clusters maintain relatively balanced distributions of 30.6% and 27.7%, respectively. These findings highlight that household energy usage patterns are not uniform throughout the year and may be influenced by seasonal changes, occupancy variations, and appliance usage behaviors. By identifying these clusters, the study enables targeted intervention strategies, optimizing resource allocation, and improving energy efficiency efforts.

The prediction models developed for each cluster demonstrate that exponential (non-linear) regression consistently outperforms linear regression in forecasting household electricity usage. The non-linear model achieves a higher average correlation (99.84%) compared to the linear model (99.67%) and yields lower error metrics across all clusters. Specifically, for the high-consumption cluster (C_2), the non-linear model achieves a Mean Absolute Percentage Error (MAPE) of 6.25%, which is crucial for accurately modeling households with higher variability in usage. These results indicate that integrating clustering with tailored non-linear prediction models can significantly enhance forecasting accuracy and better capture dynamic consumption patterns.

For future research, expanding the dataset to include additional socioeconomic and environmental variables could improve the segmentation and predictive capabilities of the model. Incorporating real-time energy monitoring data and integrating advanced machine learning techniques such as Gradient Boosting or Long Short-Term Memory (LSTM) networks could further refine predictions, especially for households with irregular consumption trends. Furthermore, exploring adaptive clustering methods that can adjust the number of clusters over time may better reflect evolving consumption behaviors. Such advancements will contribute to more precise demand forecasting, enabling energy providers and policymakers to design more effective energy management and conservation programs.

Acknowledgment

With deep humility, the author extends sincere gratitude to all parties who have offered support, encouragement, and contributions throughout the implementation of this research. This study would not have been completed without the invaluable assistance and involvement of many remarkable individuals. The author's heartfelt appreciation goes to the dedicated academic advisor, whose patience, insightful guidance, and continuous inspiration have been instrumental throughout the research process.

Gratitude is also extended to Universitas Muhammadiyah Semarang for providing the necessary facilities, opportunities, and a supportive academic environment that enabled this research to run smoothly. Special thanks are directed to the residents of RT 01, Sidorejo Village, for their openness in sharing valuable data and information. Their participation and trust played a key role in the success of this study. Finally, the author wishes to express profound thanks to family, friends, and all those whose names may not be mentioned individually—your prayers, encouragement, and unwavering support have been a source of strength at every stage of this journey. May all the kindness and support given be returned many times over.

References

[1] Z. Wei and H. Wang, "Characterizing Residential Load Patterns by Household Demographic and Socioeconomic Factors," 2021.

[2] L. Chen, Y. Cheng, Y. Zhou, Y. Zhao, and X. Liu, "Research on energy consumption in household sector: a comprehensive review based on bibliometric analysis," *Front. Energy Res.*, vol. 11, 2023.

[3] M. T. Uddina, M. Moniruzzaman, A. Selamat, and M. Y. I. Idris, "Intelligent deep learning techniques for energy consumption forecasting in smart buildings: a review," vol. 57, p. 123, 2024.

[4] M. Zekić-Sušac, T. Galović, and I. Lovrić, "Evaluating the determinants of household electricity consumption using cluster analysis," *J. Clean. Prod*, vol. 321, 2021.

[5] S. Deb, P. Ghosh, S. Dey, and D. Chakraborty, "Household electricity consumption prediction using database combinations, ensemble and hybrid modeling techniques," *Sci. Rep.*, vol. 14, p. 8598, 2024.

[6] E. A. F. R. Opoku, L. M. Abdul-Mumin, and D. Agyemang, "Forecasting household energy consumption based on lifestyle data using hybrid machine learning," *J. Electr. Syst. Inf. Technol*, vol. 10, p. 104, 2023.

[7] Rizki R. A. Siregar and B. Prayitno, "Identifying Electricity Consumption Profiles to Increase Revenue through Clustering," *J. Tek. Elektro*, vol. 10, pp. 45–52, 2021.

[8] I. K. Nti, E. O. Yeboah-Boateng, and R. K. Ameyaw, "K-means clustering of electricity consumers using time-domain features from smart meter data," *J. Electr. Syst. Inf. Technol*, vol. 10, p. 68, 2023.

[9] S. Haben, C. Singleton, and P. Grindrod, "Cluster analysis and prediction of residential peak demand profiles using occupant activity data," *Appl. Energy*, vol. 260, 2020.

[10] S. Vaiciulyte, Z. Augustaitis, and R. Zilinskas, "Analyzing and predicting residential electricity consumption using smart meter data: A copula-based approach," *Energy Build*, vol. 305, 2024.

[11] K. Wang, N. Nakagawa, and T. Nishida, "Electricity load forecasting using clustering and ARIMA model for energy management in buildings," *Jpn. Arch.*, vol. 3, pp. 65–75, 2020.

[12] J. Shen, S. Liang, H. Zhang, and P. Li, "Study on power consumption load forecast based on K-means clustering and FCM–BP model," *Energy Reports*, vol. 6, pp. 1499–1505, 2020.

[13] Ardy Herlambang, "Pengelompokkan Data Penggunaan Energi Listrik Menggunakan Algoritma Mini Batch K-means Clustering," *eProceedings Eng.*, vol. 9, pp. 1245–1252, 2022.

[14] A. Prastika, "Hubungan Antara Tingkat Konsumsi Energi Listrik dengan Pertumbuhan Ekonomi di Indonesia," *J. Ilmu Ekon. JIE*, vol. 7, pp. 18–29, 2023.

[15] G. E. Okereke, M. C. Bali, C. N. Okwueze, E. C. Ukekwe, S. C. Echezona, and C. I. Ugwu, “K-means clustering of electricity consumers using time-domain features from smart meter data,” *Journal of Electrical Systems and Information Technology*, vol. 10, no. 2, pp. 1–14, 2023, doi: 10.1186/s43067-023-00067-6.
[16] Z. Wei and H. Wang, “Characterizing residential load patterns by household demographic and socioeconomic factors,” in *Proc. 12th ACM Int. Conf. Future Energy Syst. (e-Energy)*, Virtual Event, Jun. 2021, pp. 372–376, doi: 10.1145/3447555.3465396.
[17] M. Afzalan, F. Jazizadeh, and H. Eldardiry, “Two-stage clustering of household electricity load shapes for improved temporal pattern representation,” *IEEE Access*, vol. 9, pp. 151667–151679, 2021, doi: 10.1109/ACCESS.2021.3126305.
[18] L. G. Viola, “Segmentation of household load-profiles with K-means clustering algorithm,” M.S. thesis, Politecnico di Milano, Milan, Italy, 2015.
[19] X. Jin, C. Spurlock, A. Todd, M. J. Leach, S. Scheitrum, A. Bayen, and D. Callaway, “Load shape clustering using residential smart meter data: A technical review,” Lawrence Berkeley National Laboratory, Berkeley, CA, USA, Tech. Rep. LBNL-2001095, 2017.
[20] S. McLoughlin, A. Duffy, and M. Conlon, “A clustering approach to domestic electricity load profile characterisation using smart metering data,” *Applied Energy*, vol. 141, pp. 190–199, Mar. 2015, doi: 10.1016/j.apenergy.2014.12.039.