Segmentation of Electricity Consumers Using Clustering

Martin Sarnovsky, Peter Bednar

Department of Cybernetics and Artificial Intelligence, Faculty of Electrical
Engineering and Informatics, Technical University of Košice,
Letná 9, 042 00 Košice, Slovak Republic
martin.sarnovsky@tuke.sk, peter.bednar@tuke.sk

Abstract: The work presented in this paper is focused on customer segmentation based on the electricity demand by clustering methods. The main goal was to cluster the customers of a local electrical energy distribution company into groups with similar characteristics of electricity consumption based on annual data from smart metering systems. Such customer groups can be used to ease the understanding of the differences in behaviour of individual customers, and further can be used in targeted marketing or other machine learning tasks, such as consumption prediction for specific groups. The work followed the CRISP-DM process model, a commonly used methodology for the application of data analytics in businesses. In the paper, we briefly describe each phase of the methodology and present the most important outputs. The resulting customer segments are described and interpreted, and visualizations of clusters were provided, which help to better understand the behavior of customers.

Keywords: cluster analysis; clustering, customer segmentation; energy consumption

1 Introduction

Cluster analysis is a task of unsupervised learning where the actual classification of individual objects into classes is unknown, and the values of the target variable are not provided. In general, cluster analysis involves grouping objects into clusters based on their common characteristics. Ideally, all objects within the same cluster are as similar as possible, while being distinctly different from objects in other clusters. There are various approaches to clustering, which differ based on the type of data they work with and the methods used to identify similarities.

Customer segmentation in the electricity consumption domain has evolved significantly, leveraging advanced techniques and methodologies to better understand and cater to diverse customer needs [1], [2], [3], [4]. These advanced approaches enable utilities to optimize energy management and design targeted programs that enhance customer satisfaction and engagement. Customer segmentation methods have a significant impact on customer satisfaction. By understanding the unique needs and preferences of different customer segments, utilities can offer personalized services and incentives. This tailored approach leads to higher customer satisfaction, as customers feel their specific needs are being addressed. Segmentation also allows utilities to communicate more effectively with their customers. By targeting specific segments with relevant information and offers, utilities can improve customer engagement and satisfaction. Segmentation helps design and implement programs that resonate with specific customer groups, leading to higher participation rates in initiatives related to energy efficiency and renewable energy. By leveraging segmentation data, electricity providers and distributors can enhance the overall customer experience, including providing better customer support, offering customized billing options, and addressing customer concerns more effectively. When customers feel that their provider understands and caters to their needs, it can build trust and loyalty. This positive relationship can lead to long-term customer retention and satisfaction.

The overall objective of the research presented in this paper is to develop a segmentation model for a local electricity distribution company. The main motivation for the company is to get another view on the data obtained from the users and their behavior, and utilize such knowledge in the process of tariff recommendations. Another motivation is to leverage the customer groups to create energy consumption predictive models, tailored specifically to a given customer group. This approach has the potential to be more precise than the already used general predictive models.

The paper is structured as follows. In the following section, current approaches to clustering-based segmentation techniques in the electricity consumption domain (and other related approaches) will be presented. Then, we describe the development and training of the models according to the standard methodology – we describe the data, pre-processing steps, training process, evaluation of the models, and provide an interpretation of the results followed with the proof-of-concept deployment. The final section summarizes the results and provides a brief overview of the possible future work.

2 Use of Clustering in Customer Segmentation in the Electric Energy Consumption Domain

Customer segmentation in the electricity prediction domain has seen significant advancements in recent years. With the advent of smart meters and advanced metering infrastructure, utilities can now collect detailed electricity consumption data. This data forms the backbone for customer segmentation methods. Various clustering algorithms are employed to segment customers based on their electricity consumption patterns.

Unsupervised machine learning is used in various tasks in the domain of electric energy production and distribution. Clustering methods can be used to create the demand signatures. This approach clusters customers based on their demand habits. Such segmentation can be used in several applications. Such models enable the companies to target specific customer groups for unique demand response initiatives or to design incentive programs based on customer consumption behaviours. Such advancements have enabled utilities to better understand and cater to their customers' needs, ultimately leading to more efficient and effective energy management.

In the following section, we describe some of the studies that focus on segmenting customers into clusters based on electricity consumption over a certain period of time [5] [6], [7], [8], [9], [10], [11]. Two papers present the approach that best corresponds to the aim of our work and, at the same time, provide the most comprehensive overview of the given type of issue. The first paper [12] focuses on creating intra-daily, daily, and seasonal profiles that reflect electricity consumption in households. Based on the results achieved, customers are classified into individual groups based on similar features in their behaviour. The authors of the study were provided with data from smart metering systems that recorded electricity consumption in more than 4000 Irish households in 2009 and 2010. Electricity demand, along with additional information on residents, household type, and appliances, was recorded at 30-minute intervals. The data were processed using the MATLAB programming language in combination with the statistical software SPSS. The second paper [13] focused on developing a tool that allows customers to be grouped based on meter data, which could provide electricity suppliers with a detailed overview of their customer portfolio. The analysis used data from electricity meters that recorded electricity consumption continuously in 15-minute intervals. In order to take into account the seasonal and trend components in the data, the records were transformed into a form indicating electricity consumption in different periods and seasons.

3 Customer Segmentation in a Local Energy Company

The following text will describe the overall process of training the clustering models. We proceeded with the standard, most commonly used methodology of building the data analytical models, Cross-Industry Standard Process for Data Mining (CRISP-DM) [14]. It involves several steps that should be followed to realize the data analytical task. The process starts with the Problem understanding phase. During this phase, it is important to understand the motivation why the models should be created and to translate the business needs to correct data analytical tasks, including the selection of the models and metrics, and how they should be evaluated. The following phase of Data understanding aims towards a good understanding of the available data. It means understanding the measurements, units, interpretation of categorical values, and dependencies between the features. During the Data preparation, we apply necessary pre-processing techniques to obtain the proper dataset for training of the selected models. During the Modeling phase, such models are trained, and results are evaluated. The last phase involves the deployment of the best model in the real-world environment. In this case, this step was not realized by our team [15].

3.1 Problem Understanding

During the research presented in this paper, we cooperated with a local energy distribution company, whose main activity is the electricity distribution via its distribution system to the end customer. In the area of Eastern Slovakia, the company owns a distribution system with a length of almost 21,000 kilometers. It distributes electricity to more than 600,000 households, companies, and organizations.

An analysis of customer behavior in such a large industry in the electricity sector by commonly used methods, where each customer is analyzed as a separate unit, is an extremely difficult and inefficient process. One of the main goals is, therefore, to segment customers into multiple groups (segments) based on behavioral similarities over a period of time. Analysis of such segmented customers is a much more efficient and less time-consuming process that allows the company to better understand the differences in customer behavior. Also, in addition to exploratory analysis, clustering is often used for different purposes, such as data preparation, anomaly detection, for the needs of electricity demand prediction.

3.2 Data Understanding

The dataset was provided by the local electricity distribution company in the form of two separate tables. The first dataset consisted of information about the electricity consumption of a selected sample of customers. For each customer, the data contained the time series of measurements of different variables, each recorded at 15-minute frequencies. Measurements for each customer covered the timespan of a complete year, between 01/03/2017 and 28/02/2018.

Each record in this dataset is represented by three separate rows in the table. The first two rows contain the measured values via a two-phase measurement, while the third row contains the status values of this measurement. These status values are represented in the form of a code mark defining the measured value status. The dataset has the following basic attributes:

- *Customer ID* – unique identifier of a given customer
- *Meter* - the identification of the electrometer associated with the customer
- *Variable* - value indicating the measured variable (e.g., current, voltage, power, etc.)
- *Unit* - unit of measurement (e.g., V, A, kW, etc.)
- *Timestamp* – representing the time of a given measurement
- *Status value* – represents the status of the given measurement (e.g., valid, confirmed, error, etc.)

The secondary dataset consisted of mappings of customers to electricity meter offtake points and specified the type of tariff for each offtake point.

The dataset consisted of data for 24,839 customers of the company. Measurements relevant for our study were electrical power measured in kW. Specific variables were the status variables, detailing the status of each measurement. Status variable could provide a better detailed overview, how the measurement was performed, e.g., provided by a smart meter, or inserted by a worker, or specify an error (e.g., invalid value, failure in the distribution network, etc.). The dataset contained missing values, both completely missing from the data or measurements that should be considered missing due to the status variable value. Handling of the missing values will be described in the following section.

3.3 Data Preparation

The very first step in the data pre-processing stage was to process the missing values. There were exactly 14,029,966 missing measurements in the dataset. After further investigation, the missing values were shared by 361 of the customers of a total 24 839 customers. In most cases, the missing values exceed 75% of all measured values for a given customer. Such a high number of missing values was most likely caused by the technical failure of the meters. In this case, the best option was to exclude those customers from the dataset.

The next step was to deal with status values. The status field contains values representing the circumstances of reading and recording the measurement values. It can indicate indirect measurement, estimated values, or errors that occurred in data handling. After throughout evaluation, we have selected a set of status values, which can represent erroneous or invalid measurements. We decided to consider such measurements as missing values and decided not to use them in further processing, as keeping them could negatively affect the modeling phase.

The consumption for a given customer was measured by a two-phase measurement where the electricity price changes based on low or high tariffs. That is why the electricity consumption of each customer was represented by two separate variables in the dataset. The way the measurement was realized is not important for our purposes, so we simply summed the values in these rows.

The last step of data pre-processing was an aggregation of the original dataset into several aggregated datasets used to create hourly, daily, weekly, monthly, quarterly, and seasonal behaviors. This step was taken to get multiple customer profiles based on different periods.

3.4 Models Training

For purposes described in the problem understanding phase, we decided to use partitioning methods and model-based methods of clustering. Specifically, we focused on two types of algorithms: centroid-based methods (k-means and k-medoids) [16], [17] and Self-Organizing Maps (SOM) [18]. The main idea behind the algorithm selection was to explore both parametric and non-parametric approaches (e.g., approaches where the data scientist specifies the number of clusters as a parameter of the algorithm, or it is decided by the algorithm itself).

Centroid-based methods belong to a group of algorithms that require the determination of the exact number of clusters in advance. To estimate the number of clusters during the training, we performed the elbow method (see Fig. 1). As we can notice in the chart, the sum of squared distances, which was used as a metric for cluster compactness, decreases constantly as the number of clusters grows. Therefore, the ideal number of clusters represents the point at which the evaluation metric decreases only marginally. Based on the investigation of the clusters and customer profiles, it seems that this point appears to be at a value of 5 clusters. During the training, we used 500 iterations of the k-means model and initialized the centroids using the k-Means++ method. Using the k-medoids approach, we obtained very similar results, pointing out a very similar cluster distribution with corresponding profiles. To confirm the initial clustering, we ran an experiment using a non-parametric SOM method. The method confirmed the good distribution of 5 clusters as well.

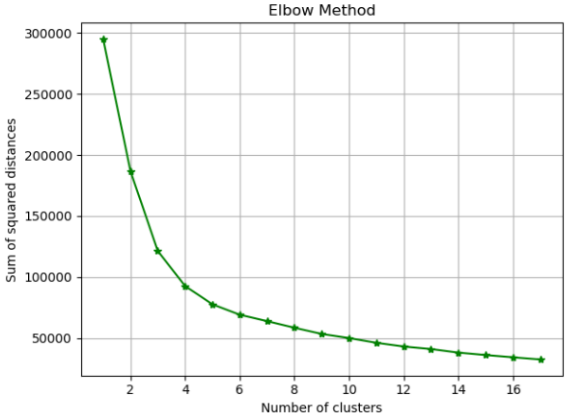

Figure 1

Elbow method application in customer clustering using the k-Means method

Tab. 1 shows the total number of customers for each cluster for each method. Following the modelling, we have performed the evaluation of the three trained models. We used the Silhouette and Davies-Bouldin metrics to compare the models and evaluate the compactness of the clusters. Tab. 2 shows the obtained metrics.

Table 1

Cluster sizes using different methods

Cluster	k-Means	k-Medoids	SOM
0	16 784	13730	16743
1	5621	4979	5696
2	1893	2985	1912
3	255	2534	200
4	37	1362	39

Table 2

Cluster sizes – clustering of the Cluster 0

Cluster	DB index	Silhouette
k-Means	0,907	0,541
k-Medoids	1,029	0,419
SOM	0,912	0,547

The differences between both metrics are minimal; however, based on a study [19] that showed that the silhouette provides an overall more accurate view of how well the clusters are formed, we decided to use the k-Means algorithm.

From the Tab. 1, it is evident that the distribution is not homogeneous, and most customers are concentrated in Cluster 1. In order to provide more in-depth segmentation, we decided to apply the hierarchical approach to clustering and decided to apply the clustering methods to this particular cluster. This largest cluster is mostly represented by ordinary households (as was investigated during evaluation). Such hierarchical segmentation could identify households with common behavior.

Household clustering was performed using the same approach as in the overall clustering, this time using only the k-Medoids method. The elbow method suggested the separation of these clusters into the four sub-clusters, with relatively improved distribution of customers (see Tab. 3).

Table 3

Cluster sizes – clustering of the Cluster 0

Cluster	k-Medoids
0	7568
1	1388
2	2326
3	5502

3.5 Evaluation

In this section, we evaluate the results achieved in terms of goals defined in the problem understanding phase. First, we had to find out which algorithm offered the best customer segmentation into individual clusters. For this purpose, we used the Silhouette score and the Davies-Bouldin index. Based on these internal validity measures, we decided to use the clusters we achieved through the k-Means algorithm.

The next step was a detailed description of individual clusters, which helps with a better understanding of customer segments whose time courses are shown in Fig. 2.

The largest cluster (Cluster 0) is mostly represented by ordinary households. Households in this cluster consume around 19,674 kW per year. We can also state that there is no difference in electricity consumption between working days and weekends. These households consume the most electricity during the winter months and at the same time the least electricity in June to September. However, this difference is not very significant.

Cluster 1 mostly consists of smaller companies that consumed 50,683 kW during the year. Electricity demand was slightly higher during weekdays compared to weekends. The curve of the monthly time courses of this cluster is almost identical to the curve of Cluster no. 0, but with more than double the electricity demand across all months.

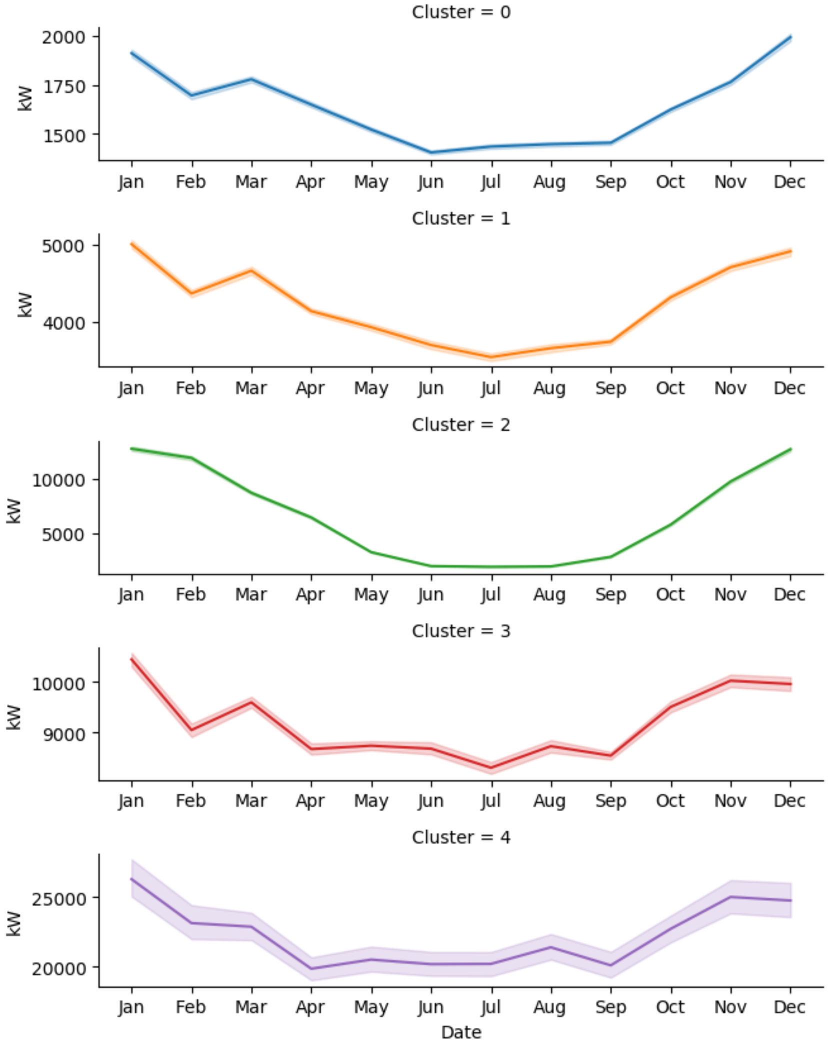

Figure 2

Yearly profiles of the particular cluster members

Cluster 2 consists of households and companies that consumed 79,759 kW during the year. Customers that belong to this cluster use electricity for heating because their electricity demand during the winter months is more than six times higher compared to summer months.

Customers in Cluster 3 consumed an average 110 184 kW during the whole year. In this case, such customers represent larger companies (or enterprises) whose electricity consumption is almost two times higher compared to other clusters (e.g., Cluster 2, which was identified as mostly consisting of smaller companies). Despite such high electricity consumption, these are not enterprises that use electricity for heating.

Cluster 4, with the smallest number of associated customers, is represented by the customers with the highest annual electricity consumption in the database (267,173 kW). As in the case of Cluster 3, companies belonging to this cluster do not use electricity for heating. Therefore, electricity consumption during the months is very similar, or differs only minimally. For the customers in this cluster, the behaviour is slightly different, as the consumption during weekdays is approximately 4 times higher than during weekends.

Interpretability of the clusters could be improved and enhanced if demographic or contextual information were integrated with consumption data. However, such data was not available at the time. More explainability of the clusters could be added using XAI (eXplainable AI) methods for time-series clustering [20], but the explanations would be similar, as no other data besides measurements were used to train the segmentation models.

3.5 Application of Clustering Results in Prediction

To demonstrate the potential use case of the segmentation, we used the cluster representatives for training the consumption prediction models. We have selected a customer for each cluster, closest to the k-Means model centroids. Those customers can be considered typical members of the given cluster. The main idea is to train the consumption prediction model for such members, which could be applied to each of the cluster members. We trained LSTM (Long Short-TermMemory) models using a lookback window of 100 days, predicting the aggregated consumption on the following day, then applied the model to predict 120 days during the year. Fig. 3 depicts the training and evaluation of the predictive model on the customer representing Cluster 3.

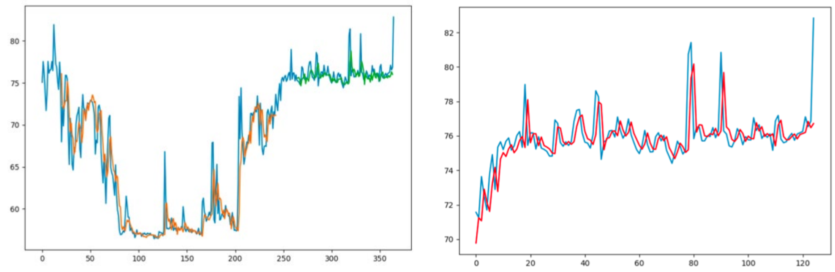

Figure 3

Left – Train (orange) and test (green) window, right – predicted (red) and ground truth (blue) values on the testing set

The models were evaluated using the MAE (Mean Absolute Error) metric. The models trained on the representatives were applied to the remaining members of the particular clusters, and we computed the average value of the MAE metric for each of the clusters. The results are summarized in Tab. 4. The results show how applicable the models trained using representatives are. While on the most frequent cluster, the model achieves relatively good results on each of members, the results on some of the other clusters (especially those, combining different types of companies) points to more diverse group of customers, with relatively higher differences in behaviour, which could lead to higher predictive error, when using this approach.

Table 4

Cluster sizes – clustering of the Cluster 0

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
MAE	4,15	40,67	58,45	24,48	49,17

3.6 Deployment

Although deployment was not a part of the study, we proposed a PoC (Proof of Concept) deployment solution for the developed models. Fig. 4 depicts the overall architecture of the solution. The most commonly used deployment scheme involves model serialization and scoring on the Python back-end REST service, with the client communicating with the API via JSON. The API receives input records, and before model scoring, it must be pre-processed in the same fashion as during the training. In the on-site deployment, this could be done directly from the database using a SQL query. First, the selected segmentation model with the selected threshold is applied to the input data and returns the cluster ID for the given record. Then, a predictive model for the particular cluster is applied (when needed). Back-end then encapsulates the response in JSON and returns it to the client. The back-end service is implemented in Python using the Flask framework to provide a REST API. For production environments, it is recommended use a WSGI server such as Gunicorn and enable HTTPS for secure communication. The server should have adequate computational resources to handle deep learning inference efficiently.

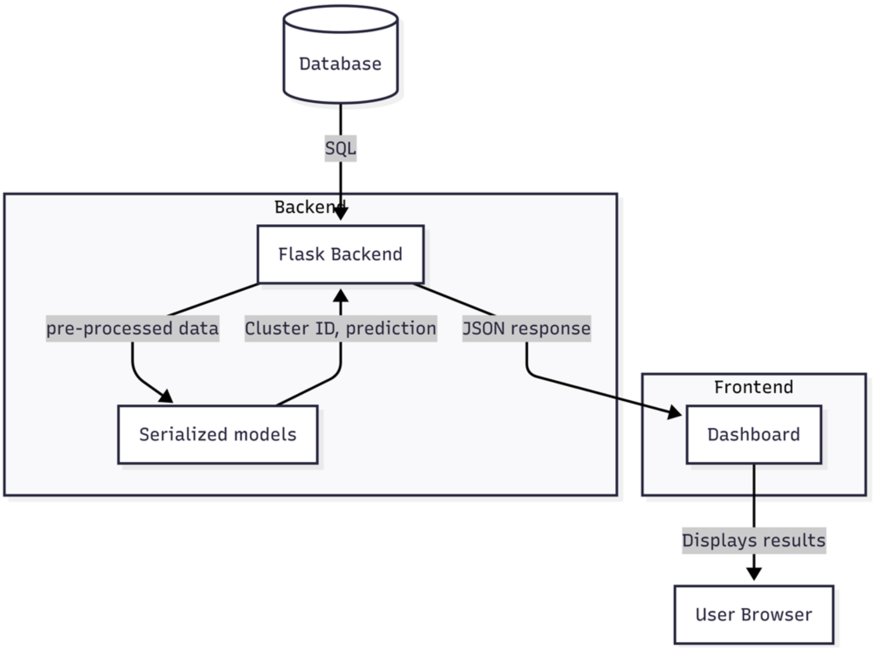

Figure 4
Architecture of the proposed deployment

Conclusions

The paper aimed to use the clustering machine learning methods to explore the customer behaviour of the electricity distribution company. The main goal was to divide the consumers into separate clusters based on their consumption profile during the year and find groups of customers with relatively similar behaviour. We used traditional machine learning approaches and, in the experiments, we compared a few selected clustering algorithms, including k-Means and Self-Organizing Maps. Models were evaluated using standard metrics, and the best model was selected for evaluation. During the evaluation, we focused on the analysis of the customers within the particular clusters. In cooperation with domain experts, we investigated the types of consumers and identified the different types of customers belonging to particular clusters. Each of the clusters is represented by the typical representative of such a group, visualized by easy-to-understand graphs, which can improve the understanding and interpretability of the given segment. Such fundamental segmentation can be then used by the company (if scaled to the entire customer database) to gain a more precise overview of the customers, to identify the possible mismatches between the customer tariffs and typical behaviour, and can lead to more precise recommendations for the tariffs for specific customers. Interesting possibility of segmentation models improvement could be in integration of demographic and more consumer-related data. Such data could help not only with interpretation of the clusters, but also to more refined and detailed hierarchical clustering. This kind of model could enable to find more specific segments on the market. More detailed clustering could also be applied in other than forecasting predictive tasks. This could involve development of anomaly detection methods, which could automatically identify possible deviations from standard customer behavior.

Acknowledgement

This work was supported by the Slovak Research and Development Agency research grant under the contract No. APVV-23-0356, NAPOLEON: Digital solutions for analysis and optimization of energy consumption.

References

[1] F. Meng, Q. Ma, Z. Liu, and X. J. Zeng, "Multiple dynamic pricing for demand response with adaptive clustering-based customer segmentation in smart grids," *Appl Energy*, Vol. 333, 2023, doi: 10.1016/j.apenergy.2022.120626

[2] F. Barjak, J. Lindeque, J. Koch, and M. Soland, "Segmenting household electricity customers with quantitative and qualitative approaches," *Renewable and Sustainable Energy Reviews*, Vol. 157, 2022, doi: 10.1016/j.rser.2021.112014

[3] H. Komatsu and O. Kimura, "Customer segmentation based on smart meter data analytics: Behavioral similarities with manual categorization for building types," *Energy Build*, Vol. 283, 2023, doi: 10.1016/j.enbuild.2023.112831

[4] T. Rahman, M. L. Othman, S. B. Mohd Noor, W. F. Binti Wan Ahmad, and M. F. Sulaima, "Methods and attributes for customer-centric dynamic electricity tariff design: A review," 2024, doi: 10.1016/j.rser.2023.114228

[5] A. Tureczek, P. S. Nielsen, and H. Madsen, "Electricity consumption clustering using smart meter data," *Energies (Basel)*, Vol. 11, No. 4, 2018, doi: 10.3390/en11040859

[6] B. McDonald, P. Pudney, and J. Rong, "Pattern recognition and segmentation of smart meter data," *ANZIAM Journal*, Vol. 54, 2014, doi: 10.21914/anziamj.v54i0.6743

[7] L. Arco, G. Casas, and A. Nowè, "Clustering methodology for smart metering data based on local and global features," in *ACM International Conference Proceeding Series*, 2017, doi: 10.1145/3109761.3158398

[8] M. Martinez-Pabon, T. Eveleigh, and B. Tanju, "Smart Meter Data Analytics for Optimal Customer Selection in Demand Response Programs," in *Energy Procedia*, 2017, doi: 10.1016/j.egypro.2016.12.128

[9] P. Lakshmanan and G. Venugopal, "Design of a Dynamic Demand Response Model Through Intelligent Clustering Algorithm Based on Load

Forecasting in Smart Grid," *Elektronika ir Elektrotechnika*, Vol. 28, No. 3, 2022, doi: 10.5755/j02.eie.30596

[10] R. K. Ahir and B. Chakraborty, "A novel cluster-specific analysis framework for demand-side management and net metering using smart meter data," *Sustainable Energy, Grids and Networks*, Vol. 31, 2022, doi: 10.1016/j.segan.2022.100771

[11] L. Wen, K. Zhou, and S. Yang, "A shape-based clustering method for pattern recognition of residential electricity consumption," *J Clean Prod*, Vol. 212, 2019, doi: 10.1016/j.jclepro.2018.12.067

[12] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Appl Energy*, Vol. 141, 2015, doi: 10.1016/j.apenergy.2014.12.039

[13] C. Flath, D. Nicolay, T. Conte, C. Van Dinther, and L. Filipova-Neumann, "Cluster analysis of smart metering data: An implementation in practice," *Business and Information Systems Engineering*, Vol. 4, No. 1, 2012, doi: 10.1007/s12599-011-0201-5

[14] C. Shearer *et al.*, "The CRISP-DM model: The New Blueprint for Data Mining," *Journal of Data Warehousing*, Vol. 5, No. 4, pp. 13-22, 2000, [Online] Available: www.spss.com%5Cnwww.dw-institute.com

[15] A. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview," *IADIS European Conference Data Mining*, No. January, pp. 182-185, 2008 [Online] available: http://recipp.ipp.pt/handle/10400.22/136

[16] M. Sarnovsky and N. Carnoka, *Distributed algorithm for text documents clustering based on k-Means approach*, Vol. 430. 2016. doi: 10.1007/978-3-319-28561-0_13

[17] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf Sci (N Y)*, Vol. 622, 2023, doi: 10.1016/j.ins.2022.11.139

[18] T. Kohonen, "The self-organizing map," *Neurocomputing*, Vol. 21, No. 1-3, 1998, doi: 10.1016/S0925-2312(98)00030-7

[19] S. Petrovic, "A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters," *11th Nordic Workshop on Secure IT-systems*, 2006

[20] Z. Huang, H. Hao, L. Du, and L. 2025 Du, "Exploring the Explainability of Time Series Clustering: A Review of Methods and Practices Exploring the Explainability of Time Series Clustering: A Review of Methods and Practices. In Proceedings 1 Topics and Motivations," *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 1005-1007, Dec. 2020, doi: 10.1145/3701551