International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 9, No. 1, 2022, pp. 39-48.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Smart Meter Data based Load Analysis Using Clustering Technique

P. Fathima Sapna [a], Dr.R. Lal Raja Singh [b]

[a] *PhD Scholar, Anna University, India. E-mail: sapnapfathima@gmail.com*
[b] *Professor, Department of Electrical and Electronics Engineering, Kalaignar Karunanidhi Institute of Technology, Coimbatore, India.*

Abstract

In this study, clustering is investigated as a method for improving energy efficiency based on smart meters for a number of household applications that are both currently available and expected to become available in the near future. These applications include smart thermostats, smart lights, smart water heaters, smart washing machines, and smart refrigerators. We describe a novel approach to load balanced clustering that is founded on the K-means Clustering algorithm. Our algorithm's major goal is to optimize network lifetime while maintaining acceptable sensing coverage in scenarios in which sensor nodes generate either uniform or non-uniform data traffic. This can be accomplished by maintaining acceptable sensing coverage. We also provide a new clustering cost function that takes into consideration not only the volume of traffic but also the amount of work that is required to communicate across substantial geographic distances. This is done so that we can achieve this objective. We demonstrate that our algorithm is able to improve both load analysis as well as load balancing in the domestic area by running extensive simulations that compare the proposed algorithm to leading state-of-the-art clustering approaches and then comparing the results to one another. This allows us to demonstrate that our algorithm is able to improve both load analysis as well as load balancing in the domestic area. In addition to this, it demonstrates the culmination of a stage in the processing of a dataset in order to compute the typical quantity of energy load that is utilized by consumers.

Keywords: Clustering Technique, Artificial Neural Networks (ANN), Support Vector Machines (SVM).

DOI: 10.9756/IAJSE/V9I1/IAJSE0918
Received: 17 December 2021/Revised: 20 January 2022/Accepted: 24 February 2022/Published: 30 March 2022

Introduction

Clustering is by far the most frequent and widely used unsupervised method for grouping data based on the similarities given by the user. It is also one of the oldest methods for doing so. It is applied in a broad variety of contexts in the games and shows that are inspired by real-life events and situations. The smart grid is one of the domains that is making use of the clustering technique for a variety of classification procedures in order to improve efficiency. Cluster analysis is the process of constructing data points by making use of the information that is obtained from the data in such a way that does not require any additional intervention from an outside source.

The information that was retrieved helps to establish a connection between the data points and guarantees that the clusters are constructed in such a way that issues that are more prevalent are clustered closer together, while issues that are less common are clustered further away from one another [1]. Researchers typically use clustering as a way for categorizing customers according to the pattern of energy usage that they exhibit in order to build an accurate customer profile. This helps the researchers produce a more accurate picture of their target audience. The classification of the customer can help to advise the plan of energy consumption to the client, which can help to decrease the amount that is owed on the bill and assist energy providers in managing peak demand.

Clustering, artificial neural networks (ANN), and support vector machines (SVM) are all elements that are incorporated into one of the methodologies for short-term energy prediction that was proposed by Torabi et al [2]. They used both historical data and meteorological data in order to get an accurate forecast of the amount of energy that will be used for the Iranian dataset. This allowed them to make an accurate prediction. Prior to the use of the prediction algorithm, the data were first categorized and divided into groups dependent on whether they were consumed during the week or on the weekends. The utilization of clustering is one way in which the accuracy of the forecast can be considerably improved. Clustering is an additional strategy that can be used to categorise customers in accordance with the pattern of energy use that they demonstrate. This work made use of the CRISP data mining approach in order to estimate the demand for short-term electrical energy.

In addition, this study generated a hybrid unique model that was built on SVM, ANN, and CBS-ANN-SVM. Both of these approaches are referred to together as CBS-ANN-SVM.

The vast majority of the time, the data that is generated by a smart meter is irregular and cannot be relied upon in any way. As a consequence of this, it is vital to fulfill the task of producing projections regarding the consumption of energy. As a potential answer to this issue, Yang et al. [3] proposed a multitask probabilistic model that is founded on Bayesian deep learning. In the next step, a method of data pooling that is based on clustering is developed in order to enhance both the variety of the data and the volume for the purpose of using the proposed framework. Because this model eliminated the problem of overfitting, not only was the accuracy of the forecast improved, but also the quality of the forecast as a whole. In the context of the smart grid, it was proved that the proposed approaches are appropriate for home demand response and network reliability study.

The amount of data that the smart meter generates is really significant. Clustering is the most effective way there is for sorting objects into the most relevant group for any study that could possibly be conducted. Using k-means clustering, Perez-Chacon and colleagues [4]) were able to establish a methodology by grouping the data from smart meters into four. The findings of the investigation were presented in a publication. An in-depth analysis of the processes that were utilized in the selection of the total number of clusters is carried out right at the beginning of the process. A variety of different indices, especially the DB-Dunn, DB-Silhouette, Davies-Bouldin, and WSSSE indices, were used to the data in an effort to make a determination on the number of clusters that exist. After taking into account the vast majority of the result in order to come at a conclusion regarding the total number of clusters, an in-depth analysis of the result is performed. The grouping of the data was accomplished according to the levels of consumption (high or low), the seasons, the days of the week, and the types of buildings (Weekdays or weekends).

Smart Metering Infrastructure System

The framework for two-way communication is made up of smart devices that make it possible for intelligent conversation to take place between different providers and consumers. When there is an issue, the sensor that was installed in the smart grid is able to identify it with a high degree of accuracy. When compared to the conventional method, the smart meter is able to collect a significant amount of additional information regarding the power use of individual customers. This reading, which is produced by the smart meter and utilized both for the creation of invoices and the monitoring of performance, On the other hand, the consumer can use it to keep tabs on their own use and exercise greater command over the quantity of electricity they draw from the grid. When putting in place the intelligent

metering component of the smart grid system, a complete method called as Smart Metering Infrastructure System (SMSI). Figure1 provides a comprehensive summary of SMSI as well as its component pieces for your perusal.

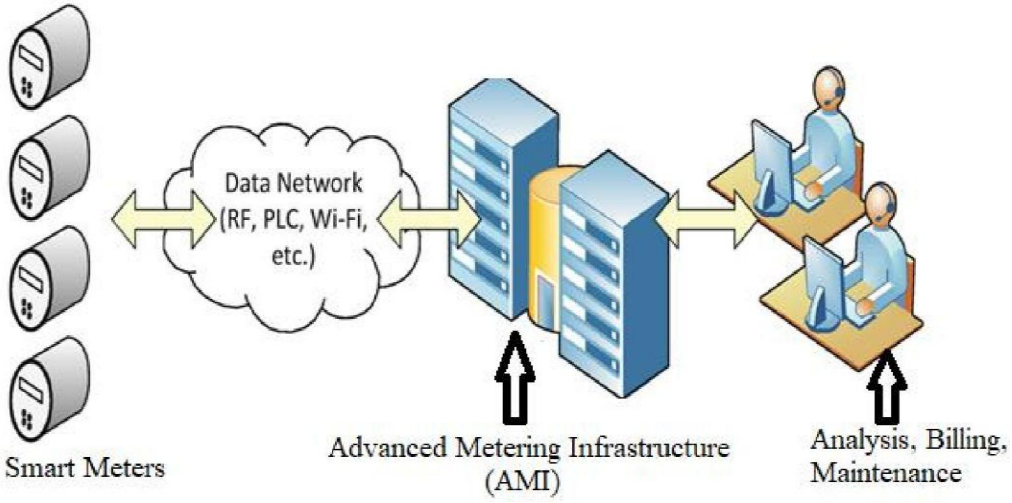


Figure 1: Smart Metering Infrastructure System

Through the processing of the data collected from SMSI-based smart meters, it is possible to determine the power consumption profile of the customer. Additionally, it is feasible to analyze future demand on the basis of present consumption data. An effective authentication system is put into place so that the information that is gathered by smart meters can maintain its authenticity while still remaining confidential. Despite the many possible benefits, SMSI has just a few number of potential drawbacks. The installation of the smart grid comes with a number of challenges, which will be broken down in the following paragraphs.

1. The SMSI's Architecture and the Components

SMSI is a collection of numerous networks and significant components that work together to perform their functions. The Smart Meter, the Collector, and the Server System, which are the Three Primary Components of the SMSI, are Managed at the Distribution Station by Someone Called the Distribution System Operator (DSO). Figure 2 illustrates the component as well as its connection to the other components in the system.

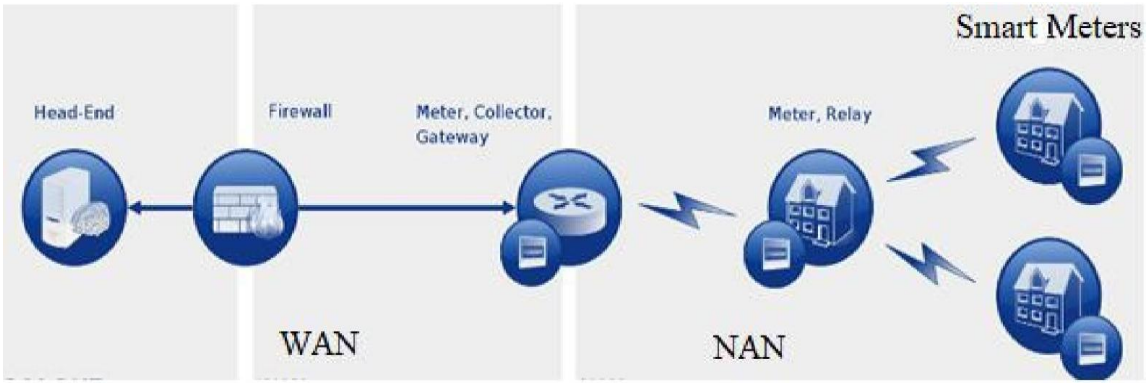


Figure 2: Smart Metering Infrastructure System & Components

The sections that follow will provide greater details regarding each individual component that makes up the SMSI.

Head-end System (HES): The head-end system refers to the metering control system that is housed within the network of an energy supply company. The HES is a software and hardware system that is responsible for receiving the streaming data in real time from the smart meters. It does this via communicating with the smart meters. In addition to this, it possesses the capacity to carry out any data validations that may be required prior to making the data accessible to other devices. After the data has been gathered, it is managed further by a metering management system that is tailored to the requirements of specific consumers.

Gateway: In the context of HES, it fulfills the roles of both a data collector and a communication node. There are some SMSI smart meters that also perform the role of gateways on their own. It serves as an interface between the home energy system (HES) and the meters that are a part of the network that is immediately surrounding it and is referred to as the neighborhood area network (NAN). During the internal communications that take place within this network, heightened security measures are in place, and both ends of this network have access to these measures.

A smart meter is one that analyzes the customer's energy consumption in real time and is installed at the customer's location. By utilizing a gateway, it is able to interface with HES in a straight-forward fashion so as to meet its needs. The smart meter is a piece of equipment that enables two-way communication and provides clients with assistance in more efficiently regulating the amount of electricity they use. Meters manufactured in today's modern times often include a user interface (UI) as well, which grants them the capability to carry out a wide range of operations. This user interface also provides instructions for installing and maintaining the device, in addition to providing a detailed handbook for the device's operation. The network that is made available within a home is referred to as a home area network, or HAN for short. Smart meters are responsible for making this network available[5].

Communication: The entirety of the SMSI is constructed out of several types of networks, all of which are utilized in order to communicate with one another. It uses a broad variety of protocol combinations at a number of distinct checkpoints in order to ensure the data and information remain private and authentic while also preventing unauthorized access to them. Regularly, SMSI relies on three important networks to carry out its business activities. In addition, there is something known as a local area network (LAN), in addition to a wide area network (WAN).

The connection that exists between the smart meter and the HES is considered to be part of a wide area network, abbreviated as WAN. Backhaul network is another name for wide area network (WAN), which is a term that is also used occasionally. The connection to the wide area network (WAN) is based on Internet Protocol (IP), and it is connected physically through a group of techniques such as fiber optic cable (FOC), digital subscriber line (DSL), general packet radio service (GPRS), multi- protocol label switching (MPLS), and power line carrier (PLC), in addition to a select number of private networks [6].

Community Network Serving the Immediate Area: It is the responsibility of the NAN to ensure that the collectors and the meters are connected to one another. It does this by establishing connections inside of the smart home between a variety of metering devices, such as those for electricity, gas, water, and heat.

A home area network (HAN) could alternatively be called a building area network (BAM) or an industrial area network (IAN), depending on the kind of user who connects to it (IAN). Through the use of the Home Area Network, or HAN, it is possible to link the intelligent home appliances that are already installed in the home to the smart grid. In order to accomplish this, it controls large HAVC (heating, ventilation, and air conditioning) equipment and regulates their output in order to keep the electricity grid steady. This makes it easier to manage use during times of peak traffic.

Clustering Techniques

Clustering is a process that organizes data into groups that are in accordance with the variables and conditions that have been assigned and include the most meaningful information. It is applied to the group in a data set that is consistent "where the within-group-object similarity is at its lowest, and the between-group-object dissimilarity is at its greatest." The subsequent sections of this paper provide more specific information regarding the procedures that were carried out in order to cluster the data that is presented in this paper.

1. Hierarchical Clustering

This approach of clustering organizes the data components into a tree-like structure and evaluates which clusters in a dataset are the most useful by comparing their levels of effectiveness. Both top-down clustering, also known as the divisive method, and bottom-up clustering, also known as the agglomerative method, are the two categories that make up hierarchical clustering. Top-down clustering is also known as the divisive method. Bottom-up clustering is also known as the agglomerative method. When using the top-down technique, sometimes referred to as the divisive method, the first step is to group all of the pieces together into a single cluster.

After that, the elements are separated into sub-clusters, and a process of recursive selection is carried out until all of the elements are positioned within the cluster that is most appropriate for them. Bottom-Up clustering, which is also referred to as the Agglomerative method, involves assigning each element to the cluster to which it already belongs, examining each element in turn, and then placing each element in the cluster that is most comparable to it.

In order to obtain an estimate of the proximity matrix before clustering, the distance that separates each point in the cluster is estimated. This allows for the production of the clusters themselves. When determining how far apart two points are, hierarchical clustering often makes use of one of four common approaches to measure the distance. Single Linkage, Complete Linkage, Average Linkage, and Ward's Minimum-Variance Method are some of the methods that fall under this category. The distance between the spots is measured in one of several distinct ways by each of these techniques. This study makes use of Ward's Minimum-Variance Method by agglomerative clustering, which is carried out with the assistance of Python programming. This method helps to reduce the amount of variation in the data. The results of the process that resulted in the creation of the dendrogram are shown in Figure 3.

2. K-means Clustering

As was said in the area of this article that was devoted to the introduction, K-means clustering is one of the unsupervised machine learning clustering approaches that has seen the most widespread application and usage. This kind of clustering has shown that the clustering of records is not suitable for usage in a number of smart grid applications [7]. The name "K" alludes to the initial selection of the number of centroids in the data set being carried out in a haphazard manner, which is the case by default. The operation of choosing the elements to the allotted centroids in an iterative manner is now being carried out.

During the process, the centroids that are the most appropriate fit for the operation are automatically assigned to ensure that the end result is of the highest possible quality. In conclusion, the outcome delivers the clusters and parts that are put together that are the ones that are the most plausible ones. It is required to have a prior understanding of the number of clusters that will act as the centroids of the data in order to use this strategy. According to equation, the purpose of the K-means approach is to get the value of an objective function known as the SEE function, which stands for the squared error function, as low as it possibly can be.

Cluster Validation

Particularly in K-mean clustering, which requires the specifications of numerous clusters to be established in advance, it is vitally crucial to determine the proper number of clusters for a dataset. K-mean clustering also demands that the suitable number of clusters be determined. Many different methods have been created throughout the years as a result of the fact that it is essential to ascertain the precise number of clusters [8]. This article examines two of the most prevalent approaches for calculating the optimal number of clusters that should be utilized in a smart grid. These methods include the elbow index [9] and the Average Silhouette index. Both of these indices can be discovered in this location [10]. Each index is judged based on its own particular profile, which is generated from the features of the individual data points. Therefore, if you want the finest results, it is essential that you utilize a range of tactics in order to achieve your goals [11].

It is possible to calculate the proper number of clusters to utilize in the categorization of the data samples via any one of a number of different methods. Some of these methods include: A few examples include the Elbow Index Method, the Average Silhouette Index Method, the Cluster Dispersion Index, and the Davies-Bouldin Index. These are only few of the methods that are available. An analysis of the relevant literature reveals that the Elbow Index Method provides greater performance when applied to data concerning power consumption. [Citation needed] [Citation needed] The mathematical equation that constitutes the execution of the Elbow Index Method is described in Section 4.1 of this document.

1. Procedure Utilizing the Elbow Index

The major goal in choosing the number of clusters to employ for any particular data set is to arrive at the optimal number that will result in the most effective grouping of the elements. This can be accomplished by identifying the optimal number of clusters to utilize for the data set. The method begins by drawing a graph with the data samples that are candidates for clustering, and then it uses the curve to determine the total number of clusters that should be utilized for the data. The methodology comprises plotting a graph with the data samples that are candidates for clustering. When K-means clustering is being done, the number of clusters should be chosen in such a way that the total Within-cluster Sum of Square (WSS) is brought down to its lowest feasible value. This should be accomplished by determining the optimal number of clusters. The comprehensive WSS takes into account both the cohesion and separation of the components that make up the cluster, which helps to guarantee that the cluster is as compact as possible.

Data Summary & Preparation

In light of the fact that the prior part discussed all of the strategies and methods that were utilized in the process of calculating the energy load profile, we will now discuss the information that is related to the data set. The data for this study came from the power department in Himachal Pradesh, India, and it was used in the investigation that was carried out. Sixty-five records of the information that was gathered by the smart meter on a daily basis regarding the amount of electricity that was consumed are included in the data samples. The information obtained from the meters is classified as commercial because they are all situated within the business district. The reading is captured at consistent intervals of thirty minutes during the entire process.

In order to gain a better understanding of the behavior of customers, each measurement is a load curve that is based on a reading taken over the course of 24 hours. The fields that were included in the records were the meter number, the name of the customer, a timestamp with a 30 minute interval, and the total quantity of energy use expressed in both KVA and KW units. The data collection took place over the course of a span of six months, commencing in July 2020 and extending all the way through December 2021. It was determined that a total of.56 million representative samples should be taken from the data.

The purpose of this paper is to apply a clustering approach to the data that is available and to organize the data into the most relevant cluster that can be created from the data. After that, these clusters are studied in order to ascertain an average energy consumption and the energy usage of the majority of users. This is done so that the data may be used to make informed decisions. By providing information that is helpful, this study can make a contribution to the prediction of the load forecasts in later studies. The data is then processed and cleaned up with the help of the analytic tools that are offered by IBM SPSS in order to get ready for further analysis.

The statistics allow for the identification of the consumption patterns that are exhibited by the companies. Following the processing of this data set, the statistics for the daily energy consumption, average load, and average daily consumption may be generated for each cluster. Following the processes of cleaning, identifying missing values, and deleting outliers, we were left with a total of 65 meter records, despite the fact that the initial total number of meters was 69. As was said before, we make use of the Elbow Index in order to figure out how many clusters should be used to provide the best possible results. Figure 3 displays the results of the experiment. When reviewing over all of the data that was accessible, it was found that the majority of researchers supported utilizing the Elbow Index Method when deciding how to proceed with the electricity cluster.

Using Python and the Elbow method, we came to the conclusion that the ideal amount of data clusters to build was four, and we utilized Python to create these clusters. Before deciding to settle on the current configuration of four clusters, we first attempted to partition the data into five and six clusters. When we took five clusters, we got one cluster that had just one record, and when we took six clusters, we got one cluster that contained no elements at all. However, when we took five clusters, we received one cluster that included only one record. In summary, the data samples have been partitioned into the aforementioned four separate categories.

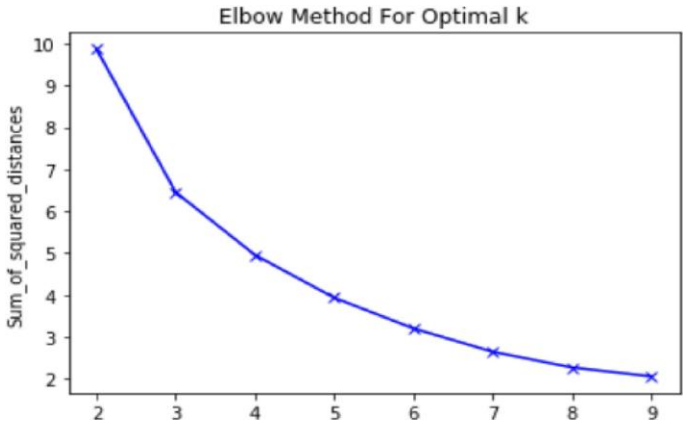


Figure 3: Optimal Cluster Validation – Elbow Method

Before we arrive at a conclusive answer on the optimum number of clusters, we put our skills at making five and six clusters to the test. When we first began, there were five clusters, but only one of them produced a value for us. After increasing the number of clusters to six, we discovered that one of the clusters did not contain any values. At long last, the information has been separated into its four discrete categories.

Results & Discussion

Python is used to implement the agglomerative clustering approach, which is used to organize the data samples into clusters. Python is also used to do the organization of the data samples. After that, the dataset is evaluated with the help of the analytical tools that are included in Microsoft Excel in order to compute the total energy usage (KW) for the past six months, and the findings are then averaged. As was previously discussed, the Elbow index approach advocated segmenting the information that was gathered into four separate clusters. Figure 4, which can be found below, displays the dendrogram that was generated as a result of using Python as a programming language. The amount of energy that was consumed may be viewed along the y-axis, while the number of meters that were covered can be seen along the x-axis. The graphic illustrates in a clear and concise manner four separate groups that are arranged in a hierarchical approach.

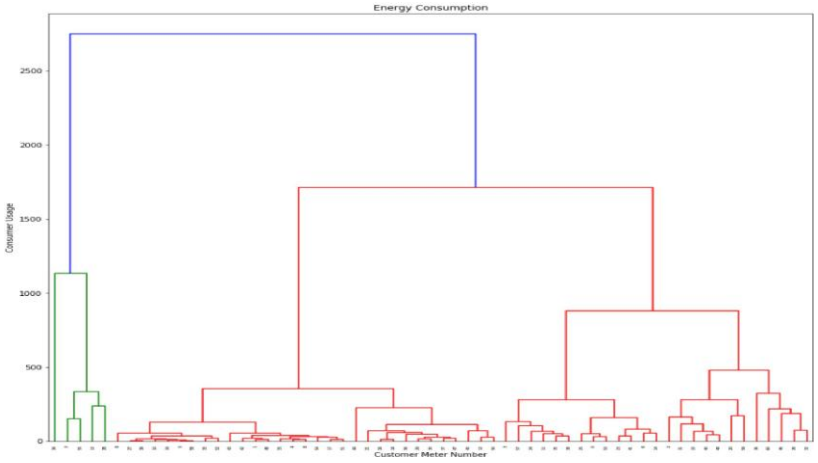


Figure 4: Dendrogram based on Average Power Consumption in KW

The process of extracting elements from clusters is accomplished with the help of K-Means clustering, which is carried out in Python. Table-1 provides specific information regarding the typical amount of power that one component of the specified cluster used up over the course of a period of half a year. Before the data was cleansed, there were 69 total customers, but once it was processed, there were only 65. The K-means algorithm is utilized, and Python Sklearn is the tool of choice for performing the clustering of the data. It does not require any action on the part of the user in order to organize the data in the most pertinent way. The following figure demonstrates how the clustering of the data using K-Means led to the following results.

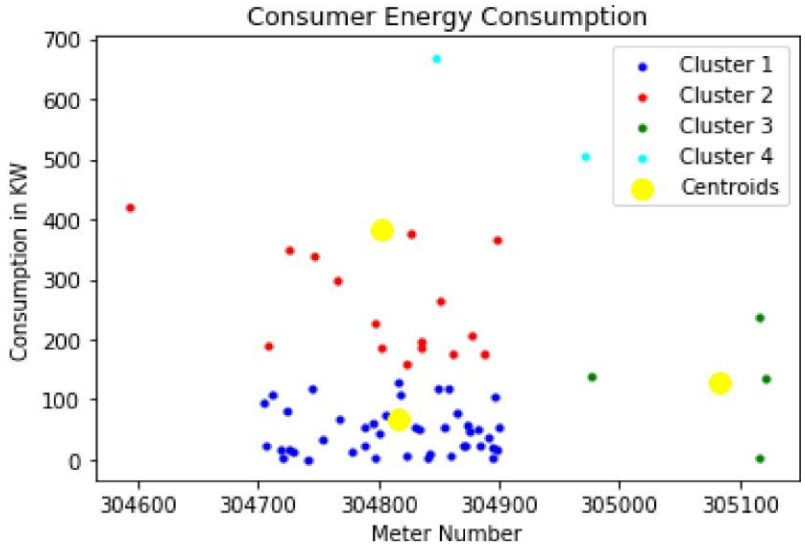


Figure 5: Demonstrates how the Elements are Distributed throughout each Cluster

This chart can be accessed here. An explanation of the distribution of the total number of customers in the cluster that is most suitable to them may be found in table 1 below.

Table 1: Details of Clusters Formed on the Complete Dataset

Cluster	Number of Consumers	Average Energy Consumption (KW)	Average Daily Consumption (KW)
C1	14	100.17	0.04332
C2	33	27.88	0.01212
C3	5	411.90	0.15320
C4	13	191.17	0.07212

The load profiles for each cluster are displayed using a scattered chart so that it is easier to get a clearer picture of how much consumption the consumers in each cluster are generating. The table that follows gives an overview of the components that make up each group, including the number of different components that are present, the average consumption over the course of half a year, and the average consumption on a daily basis.

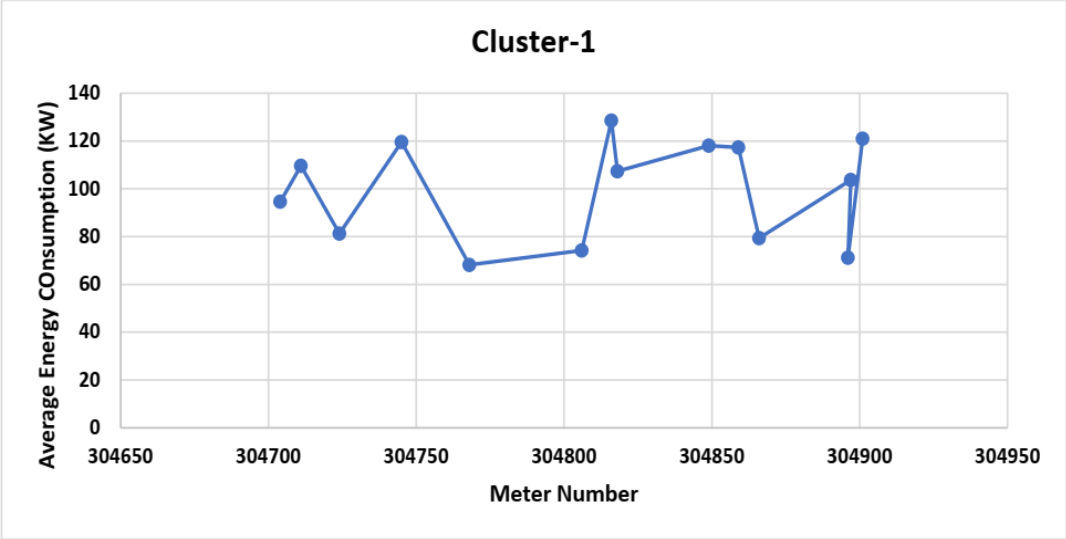


Figure 6(a): Average Consumption of Energy in the Cluster-1

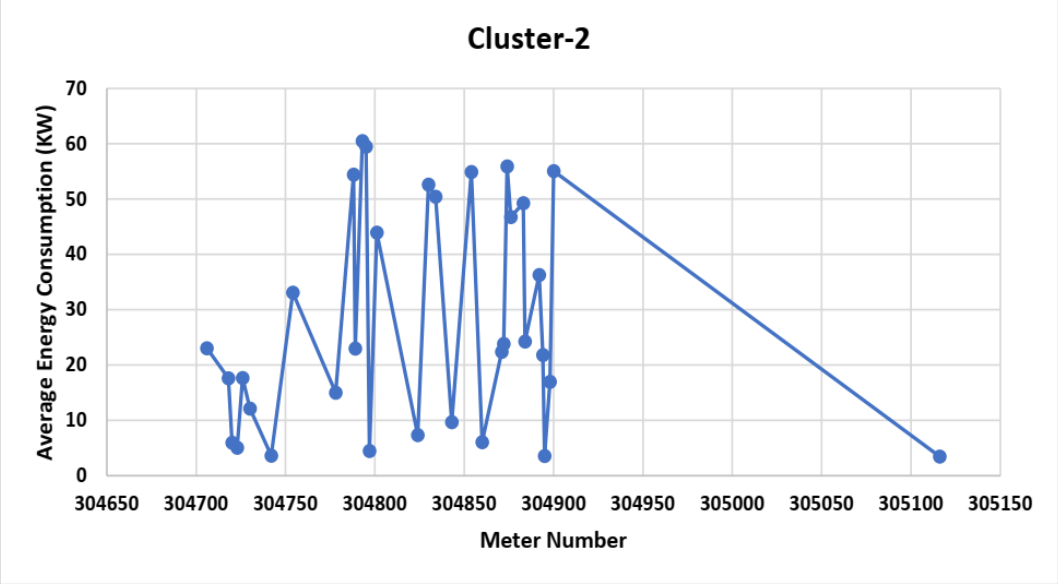


Figure 6(b): Average Consumption of Energy in the Cluster-2

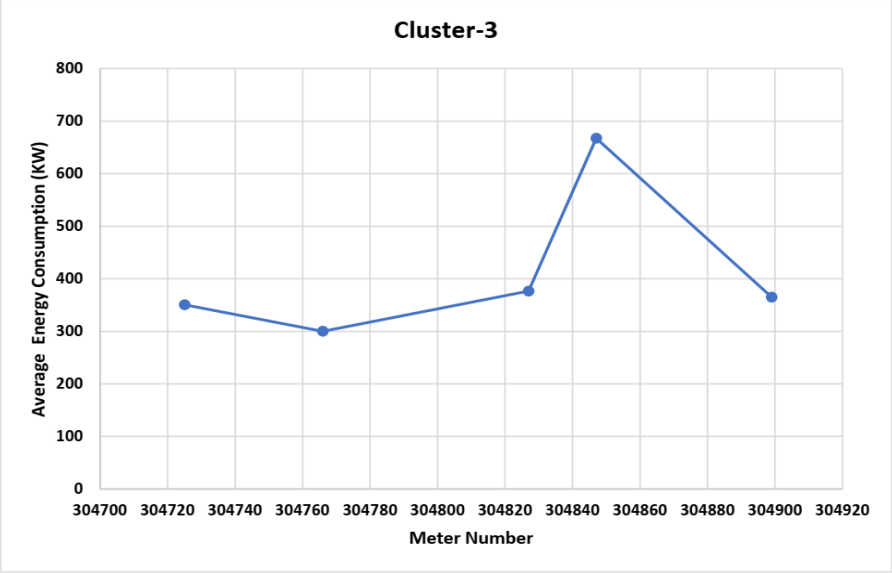


Figure 6(c): Average Consumption of Energy in the Cluster-3

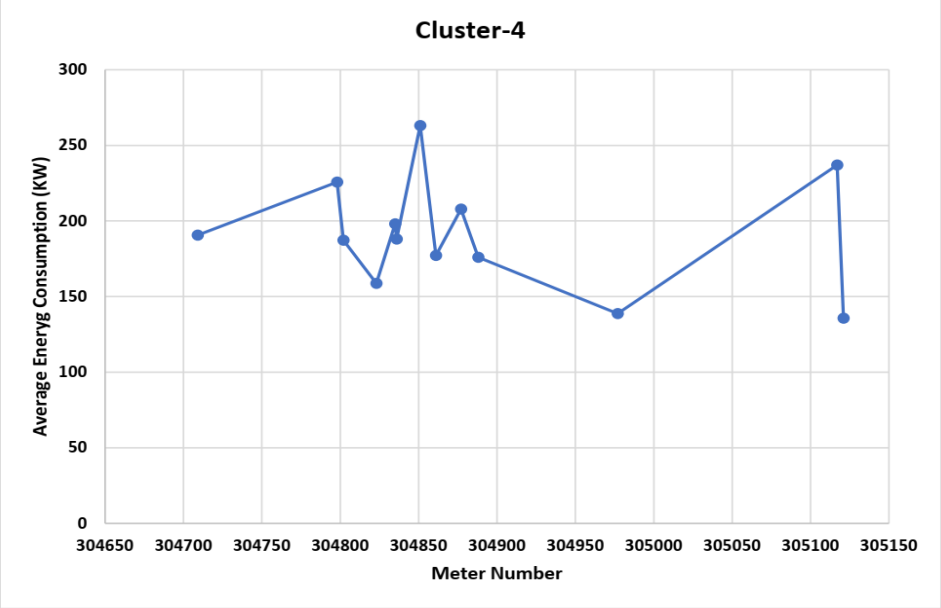


Figure 6(d): Average Consumption of Energy in the Cluster-4

This finding contributes to the classification of various business sectors according to the energy load that they are responsible for consuming on average. We are able to make further use of this information to estimate the peak load for each organization and prepare measures to optimize energy usage during times of peak load. This will allow us to save money and reduce our environmental impact. The processing of the long-term energy demand prediction will also benefit from this outcome thanks to the usefulness of the outcome. Values 6 (a-d) show the quantity of energy that was used by each individual customer inside their respective clusters. These figures represent the total amount of energy that was utilized.

Clustering customers according to their energy consumption and comparing that consumption to the consumption of other customers is extremely helpful when making a variety of different planning decisions, as the previous visualization demonstrates. These decisions include designing tariff structures, developing optimal energy plans, implementing demand-side management, and making a variety of other planning decisions. The fact that four distinct facets of the business are depending on the load consumption profile was just illustrated through the use of graphical representations of the data.

Cluster 1 has an average energy consumption that falls somewhere between 65 and 130 kW, Cluster 2 has capacities that range anywhere from 1 to 60 kWh, Cluster 3 has a range that goes anywhere from 300 to 650, and Cluster 4 has a range that falls anywhere from 70 to 250 kWh. The cluster-2 average minimum load category accounts for the vast majority of the overall majority usage rate, as shown in figure 6(b), which may be viewed here. The cluster-3 type is the intermediate maximum type. A representation of this type may be found in image 6. In addition to that, this cluster is comprised of a variety of businesses operating in a variety of different sectors.

Conclusion

This paper will take you step-by-step through an entire step in the process of processing a dataset in order to calculate the average amount of energy load that is consumed by customers. In addition to this, it gives an idea of the most common methods that are used for estimating the number of clusters for any data set and the types of clustering techniques that are used for data mining. In other words, it provides an overall picture of both of these aspects. Python is used for each step of the process, and the process is carried out while taking into account a data set on the amount of energy consumed. The processing of the raw data is carried out with the assistance of a piece of software known as IBM SPSS. For the purpose of getting a deeper comprehension of the topic at hand, the results are presented in the form of a dispersed graph for easier visualization. We intend to make use of this result in the future to create predictions regarding the energy demand, and we also intend to optimize energy usage during peak load hours by providing enhanced energy policies and incentives. In the future, we expect to make use of this result.

References

Singh, S., & Yassine, A. (2018). Big data mining of energy time series for behavioral analytics and energy consumption forecasting. *Energies, 11*(2), 452.

Torabi, M., Hashemi, S., Saybani, M.R., Shamshirband, S., & Mosavi, A. (2019). A hybrid clustering and classification technique for forecasting short-term energy consumption. *Environmental Progress & Sustainable Energy, 38*(1), 66-76.

Yang, Y., Li, W., Gulliver, T.A., & Li, S. (2019). Bayesian deep learning-based probabilistic load forecasting in smart grids. *IEEE Transactions on Industrial Informatics, 16*(7), 4703- 4713.

Pérez-Chacón, R., Luna-Romera, J.M., Troncoso, A., Martínez-Álvarez, F., & Riquelme, J.C. (2018). Big data analytics for discovering electricity consumption patterns in smart cities. *Energies, 11*(3), 683.

Bennett, C., & Highfill, D. (2008). Networking AMI smart meters. *Paper presented at the 2008 IEEE Energy 2030 Conference,* 1-8.

Rafiei, M., & Eftekhari, S.M. (2012). A practical smart metering using combination of power line communication (PLC) and WiFi protocols. *Paper presented at the Proceedings of 17th Conference on Electrical Power Distribution,* 1-5.

Viegas, J.L., Vieira, S.M., Sousa, J.M., Melicio, R., & Mendes, V. (2015). Electricity demand profile prediction based on household characteristics. *Paper presented at the 2015 12th International Conference on the European Energy Market (EEM)*, 1-5.

Granell, R., Axon, C.J., & Wallom, D.C. (2014). Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles. *IEEE Transactions on Power Systems, 30*(6), 3217-3224.

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications, 105*(9).

Ferreira, A.M., Cavalcante, C.A., Fontes, C.H., & Marambio, J.E. (2013). A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector. *International Journal of Electrical Power & Energy Systems, 53*, 824-831.

Tureczek, A.M., Nielsen, P.S., Madsen, H., & Brun, A. (2019). Clustering district heat exchange stations using smart meter consumption data. *Energy and Buildings, 182*, 144- 158.