POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Ingegneria Elettrica

PERFORMANCE ASSESSMENT OF LOAD PROFILES CLUSTERING METHODS BASED ON SILHOUETTE ANALYSIS

Relatore: Prof. Alberto Berizzi

Correlatore: Dott. Alessandro Bosisio

Tesi di Laurea Magistrale di:
Holguer H. Noriega Zambrano
Matr. 10642448

Academic Year 2019-2020

ACKNOWLEDGEMENTS

Special thanks to God and my Princess Maria for giving me the opportunity to fulfill my dream of studying in a worldwide top university. Gratitude to my parents, Paula and Holguer, to my sister Paola for supporting me every second before and during this trip. I also thank my aunt Mercedes, my aunt Sonia, my aunt Juana and all those people who helped me during my stay outside my country.

Thanks to all my master's professors who shared their knowledge with me, especially Professor Berizzi and PhD. Alessandro for the opportunity to do the thesis in their research department, and for their constant support for the development of this thesis.

A special recognition to the state of Ecuador for its support during these two years of master's degree, through the scholarship program of SENESCYT that began in 2007, and that for more than 10 years has helped thousands of Ecuadorians to study in the best universities in the world.

Finally, the work committed to this thesis is offered to my country Ecuador, to my beloved city Guayaquil- which was hit hard by COVID-19, my uncle Lupercio, my uncle Wacho, my uncle Wester, my cousin Rolandito, my best friend Jorgge and my grandfather Carlos who from somewhere out of this world are watching me raise the name of my country.

INDEX

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Abstract in Italian

Questa tesi tratta una metodologia per raggruppare i profili di carico di utenti di MT sulla base di algoritmi di clustering e tecniche di cluster validation. Per l'analisi sono stati utilizzati dati reali acquisiti ogni 15 minuti del distributore di energia elettrica e gas naturale di Milano UNARETI. Lo scopo di questo studio è analizzare la bontà di alcuni metodi di clustering basandosi principalmente su approcci di cluster validation.

La metodologia di clustering include le seguenti fasi: caricamento dei dati di carico dal database sorgente; filtraggio e semplificazione dei dati; analisi di alcuni campioni rappresentativi e successiva analisi dell'intero database; analisi di sensitività basata sulla modifica dei parametri di input dei metodi di clustering utilizzati; identificazione della migliore rappresentazione dei dati e del miglior time-step; analisi del valore della funzione Silhouette e valutazione dell'efficacia del clustering utilizzando il valore medio della funzione Silhouette e grafici di Elbow; normalizzazione delle curve per i diversi metodi di raggruppamento.

Sono state analizzate tre tipi di rappresentazioni: una rappresentazione in termini di valore assoluto utilizzando direttamente i dati acquisiti dal campo e due rappresentazioni di valori in per unità. Inoltre, sono state analizzate tre diverse fasi temporali: 15 minuti, ora e giorno. Tra questi è stata scelta la migliore combinazione che massimizza i valori della funzione di Silhouette. La metodologia è stata prima testata su campioni e poi replicata sull'intero database.

Saranno considerati cinque metodi di clustering: DBSCAN, Hierarchical, K-Nearest Neighbor, K-Mean. I valori della funzione di Silhouette verranno utilizzati per testare le prestazioni di ciascun metodo. Sono stati utilizzati i grafici dei valori della funzione di silhouette e i grafici di Elbow.

Analizzando i risultati è stato possibile trarre le seguenti conclusioni: i migliori risultati sono stati ottenuti quando l'analisi ha lavorato con la rappresentazione di valori assoluti e un intervallo di tempo di 15 minuti; nella maggior parte dei metodi utilizzati la quasi totalità delle curve di carico sono associate ad un unico grande cluster.

Abstract in inglese

The thesis is about a methodology for grouping MV substation load profiles based on clustering methods and cluster validation techniques. Real data from the DSO UNARETI was taken each 15 minutes for one year. The aim of this study is to analyze the performance of some clustering methods mainly based on cluster validation approaches.

The clustering methodology includes the following steps: Loading raw load data from the database; load data filter and load data simplification; analysis of samples and subsequent analysis of the entire database; sensitivity analysis based on changing the input parameters; best data representation and best time-step using the samples; Silhouette value analysis and clustering validation using average silhouette value and Elbow charts; normalization of curves for the different clustering methods.

Three types of representations were analyzed: the representation of absolute values obtained directly from the measurements and two representations in per unit values. Moreover, three different time steps were analyzed: 15 minutes time step, hourly and daily time step. Among them, the best combination that maximizes the silhouette values was chosen. The methodology was first tested on samples and then replicated to the entire database.

Five clustering methods are considered: DBSCAN, Hierarchical, K-Nearest Neighbor, K-Mean. The silhouette and clustering validation values will be used to test the performance of each method. The use of graphs of silhouette values and Elbow charts were used.

By analyzing the results of the load clustering, the following conclusions could be drawn: the best results were obtained when the analysis worked with absolute values representation and a time-step of 15 minutes; in most of the methods evaluated there were a tendency of having a big cluster which counts for almost all the input load profiles.

1 INTRODUCTION

One of the pillars in Power System is to know what the variation in the electrical load through time is. The graphical view of this variation called load profile is used by Power producers to make plans about how much electrical energy should be available in the future. The load profile will change according to many factors like holiday seasons, temperature, customer infrastructure, etc. In this study, the methodology analyzes four different clustering methods and their variations are proposed and based on the average Silhouette value, the method with the best performance will be chosen. The methodology uses real data from the electrical substations of the Italian distribution company UNARETI, and the load profiles are analyzed in detail.

1.1 RESEARCH BACKGROUND

Every year in the world, the load grows rapidly. Thus, an analysis of power load characteristic is a crucial and useful factor for IDMS (Integrated Database Management System) used by DSO (Distribution System Operator) for planning and managing the Distribution Network.

IDMS is an example of state-of-the-art planning technology; it gives engineers the opportunity to do remote planning and do real time simulations without affecting the normal Distribution operation.

In order to know the quantity of load it is necessary to get information using meters and later use some statistical methods We can know about the *actual demand from meter devices at strategic locations in substations; this is beneficial to both distribution and end-user.* Furthermore, based on historical measurements of power; it is possible to analyze the power load data for future uses, using methods like clustering, disaggregation, forecasting. To do the above, it is necessary to apply accurate load profile methods using simulations, beneficial to both the power system and users.

From the power system side, the application of these methods can increase the reliability and stability of power operation, enhance the quality of power dispatch, do better public and private investment in network infrastructure, improve the efficiency of power generation, reduce fuel consumption, etcetera. These load profile methods will provide

authorities better guidelines for formulating policies about production and operation in the system.

From the user's side, the load profile methods will offer a reduction in electricity bills, which reduces production costs in the industrial sector and helps to the urban and rural home economies. The operation is based on move the electricity consumption during peak hours to valley hours, better distributing the load on the system, give an idea to the factories and citizen how to plan better.

1.2 LITERAL REVIEW

There is a lot of information on clustering methods in the scientific literature, many of them applied to various fields such as medicine, marine life, etcetera. The advantage of these clustering methods is their mathematical basis, since it allows their replication in various areas, such as in our study of power systems.

In recent years, researchers have proposed a variety of clustering methods. Various methods for clustering load curves have been used in the load clustering in recent years such as K-means Hierarchical methods, K-Nearest Neighbor, DBSCAN methods, etcetera. Some of them can be used at the same time or together.

In the data mining technologies, instead of improving the method of clustering, decreasing the dimension of the data and extract the main features of the load profile can be another important method to cluster [1].

Many of these methods can be built using computer software such as MATLAB or Python. There is a variety of documentation in which researchers can consult about the different codes implemented in the past, and which could be improved and adapted to the need of the present problem.

1.3 THESIS CONTENT

This thesis analyzes the daily load data of more than 1500 MV customer substations in Milan over a 15 minutes step measurement for 365 days. The procedure to analyze the data is a clustering methodology.

Based on the basic theory of clustering of load curves, several processes were carried out within the methodology:

a. First a filter of the raw data was made, using Microsoft Access and MATLAB. Later, the data pertaining to Milano was separated. Lastly, a PCA (Principal Component Analysis) process is used to reduce the dimension of the whole data in order to ease the computations.
b. Select the best data representation and time-step. A sample analysis could be used to do it. To improve the analysis, it is recommended to use some set of samples chosen at random. The conclusion obtained in the analysis of the samples could be used for the entire database.
c. To select the best clustering method and its adequate number of groups from those we are analyzing, the silhouette values will be used, and their average value per method. Each method is characterized by input parameters, which can give different final silhouette values, then a sensitivity analysis will be carried out in order to get the best combination of input data.
d. Once the best combination of data representation and time-step is known, the sensitivity analysis will be applied again to the entire database. In this way, two cluster validation processes will be applied: Elbow charts and Average Silhouette Value.
e. The Elbow and the Average Silhouette Value approach will be used to make comparisons between methods. In the sensitivity analysis, those single-member groups with silhouette values equal to 1 are not taken into account.

The thesis is organized as follows:

a. The basic concepts and theories of clustering methods. This chapter mainly introduces the different clustering methods that exist, what are their advantages, their drawbacks, and the mathematical basis of each one.

b. The basic concepts and theories of Silhouette value methodology. This chapter mainly introduces the basic principles of the silhouette value, the mathematic basis and how is the connection of this score with the different methods, through the input parameters.
c. The methodology to be used to study the best clustering method and the appropriate number of clusters. This chapter introduces the four steps of the procedure: treatment and filter of the raw database, sensitivity analysis of the sample and the entire database, selection of the best data representation and time-step of the samples and subsequent application to the entire database and finally, the comparison between the chosen methods and subsequent selection of the best clustering method with its adequate number of clusters.
d. The result analysis. In this chapter, each step of the methodology process is shown through graphs and tables. Finally, the selection criteria for the selection of the appropriated clustering method is shown based in the cluster validation approach.
e. The observations conclusions of the study will be shown in a summarized form.

2 BASIC CONCEPTS OF CLUSTERING

2.1 Definition of clustering

In the world we live, every day people deal with different kind of information like measurements and observations. The information provides us the basis for analysis, decisions, and understanding of all kinds of phenome in nature. We can infer the properties of a specific object based on the category to which it belongs.

Basically, classification systems are either supervised or unsupervised, depending on whether they assign new data objects to one of a finite number of discrete supervised classes or unsupervised categories, respectively [2].

The methodology to be use is the unsupervised classification, also called clustering. Clustering is the ability to lump together objects with similar characteristics. There is not universally agreement about the definition of the term cluster. The main idea of clustering procedure is to group a set of N elements in K clusters.

The goals of cluster analysis in the following four major aspects [3]:

• Development of a classification.

• Investigation of useful conceptual schemes for grouping entities.

• Hypothesis generation through data exploration.

• Hypothesis testing or the attempt to determine if types defined through

other procedures are in fact present in a data set.

There are 3 main reasons of doing clustering:

- A good clustering method has a great predictive potential. Researchers try to create groups of their data because they want a better description of it, as this improves decision-making
- Cluster gives the possibility to compress the information into smaller parts, for example the center of the clusters, i.e., centroid.

- Other reason of clustering is when the clusters are made, they help us to identify the "outliers", i.e., the cases in which clusters fail to accurately represent data…

2.2 Procedure of cluster analysis.

Feature selection chooses distinguishing features from a set of candidates, while feature extraction utilizes some transformations to generate useful and novel features from the original ones [4].

Clustering algorithm design or selection: This step is about to find a mathematical and graphical method to determine a proximity measure and like this build a criterion function. Once a proximity measure is gotten, clustering could be built as an optimization problem.

Cluster validation: Evaluations after the clustering algorithms based on standards and criteria are important to give users a degree of reliability for the clustering results.

Result interpretation: This last step is about giving users useful interpretations from the original data thus they can develop a clear understanding about the data and then take decisions [5].

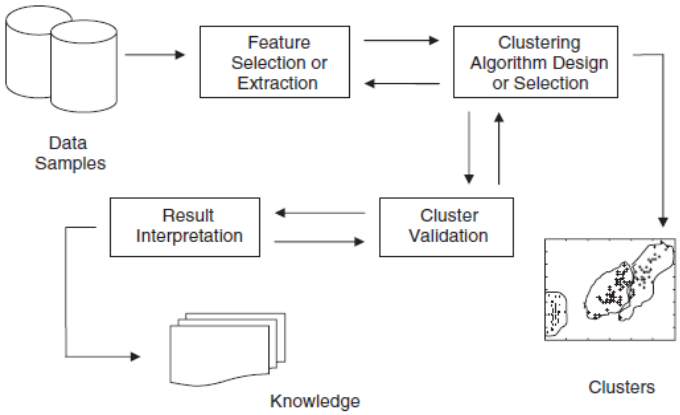


Figure 2.1 Clustering procedure. The basic process of cluster analysis consists of four steps with a feedback pathway [4].

2.3 Requirements of clustering

There are some requirements to take in account about clustering algorithm [7]:

- Trading with different types of data properties.
- Scalability
- Finding clusters with arbitrary and randomly structure.
- Capability to handle noise and outliers in the database.
- insensitivity to order of input records.
- High dimensionality.
- Interpretability and usability.

2.4 Problems about clustering approach

There are several problems with clustering [7]:

- Current clustering techniques do not address all the requirements adequately (and concurrently).
- Trading with large number of dimensions and large number of data items can be problematic because of time complexity.
- The effectiveness of the method depends on the definition of "distance".
- When a distance does not exist, we must "define" it, but it is not always easy, especially in multi-dimensional spaces.
- If we do not have the enough background about the problem to be researched, the result of the clustering algorithm can be understood in different ways.

2.5 Uses of clustering

Clustering has been applied in a wide variety of fields, as illustrated below with several typical applications [5].

1. Engineering

2. Computer sciences
3. Life and medical sciences
4. Astronomy and earth sciences
5. Social sciences
6. Economics.
7. And some other topics.

2.6 Clustering Methods

The clustering methods considered in this study are as following:

K-MEAN METHOD

K-means is one of the basic unsupervised learning algorithms that fit and solve some clustering problems in a very good way. The technique organizes in a simple and easy way a given database using a specific number of clusters decided before starts the clustering.

The principal idea is to define k centroids, one for each cluster. The locations of the centroids should be in a smart way in order to get better results, for example placing them far away from each other [10].

The steps to perform the K-MEAN are as follows:

- Place the K centroids far away from each other.
- Associate each point in the database to the closest centroid. When all the points are associated, the first part of this step is complete, and a group is formed.
- A recalculation of the k new centroids should be done as barycenters of the clusters resulting from the previous step.
- After the K new centroids are gotten, a new attachment has to be done between the same data set points and the nearest new centroid. A loop has been generated.
- The outcome of this loop says us that the k centroids change their location step by step until no more changes are done.
- The goal of this algorithm is to minimize an *objective function*, in this case a squared error function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| X_i^{(j)} - c_j \right\|^2 \quad (2.1)$$

Where:

- J is an indicator of the distance of the n points from their respective cluster centers.
- $\left\| X_i^{(j)} - c_j \right\|^2$ is the distance between a point $X_i^{(j)}$ and the center of the cluster c_j.

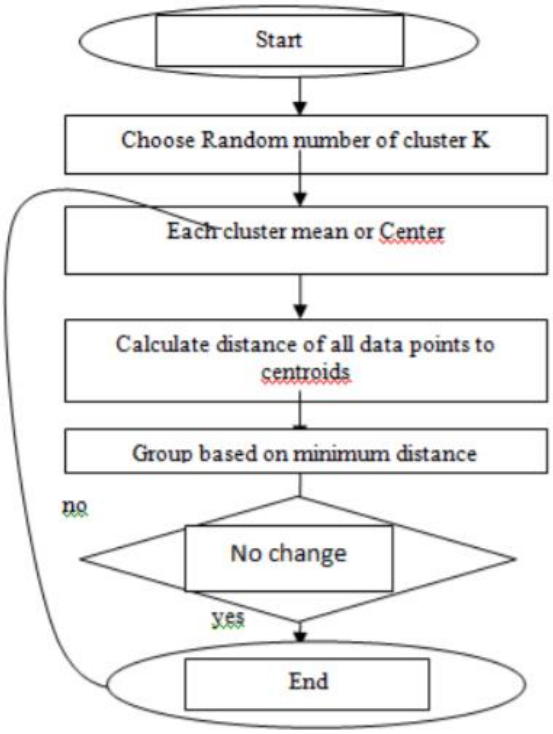


Figure 2.2 K-Mean Clustering Process [11].

There are many advantages to using K-Mean, among which are:

a. The algorithm is very sensitive to the initial randomly selected cluster centers.
b. Additionally, K-Means is often more suitable than Hierarchical clustering for large amounts of data.
c. Fast for low dimensional data.
d. If the number of clusters is large, the K-Mean method can fine purely sub clusters.
e. It is simple to codification.

On the other hand, there are also some limitations to using K-Mean, among which are:

a. K-Means has problems when clusters are of differing of sizes, densities, non-globular shapes.
b. It required to know in advance how many clusters the problem has.
c. Not suitable to find clusters with an arbitrary form.

HIERARCHICAL CLUSTERING

Hierarchical clustering method clusters data using a range of different scales by creating a dendrogram (a cluster tree). The elements that were viewed as most similar by the members in the study are placed on branches that are close together [11].

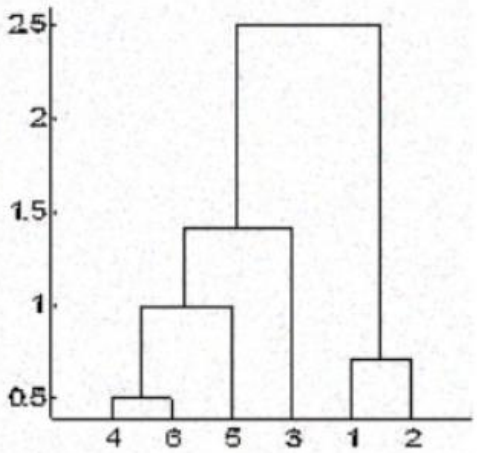


Figure 2.3 An example of hierarchical tree [11].

The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level combine to form clusters at the next level. This multilevel hierarchy allows you to choose the level, or scale, of clustering that is most appropriate for your application [12].

Dendrogram

In Fig. 2.3 , elements 4 and 6 are combined into one cluster, let say group 1, since they were the closest in distance. Elements 1 and 2 will be grouped in cluster 2. Element 5 was joined in the same group 1 followed by element 3 resulting in two clusters. At last the two clusters are merged into a single cluster and this is here the clustering process ends.

The criterion used to stop the method has a lot to do with the prior knowledge of the data. But sometimes we do not have the complete information. In such cases, you can control the results from the dendrogram to approximate the number of clusters. You cut the dendrogram tree with a horizontal line at a height where the line can traverse the maximum distance up and down without intersecting the blending point. In the above case it would be between heights 1.5 and 2.5 as shown:

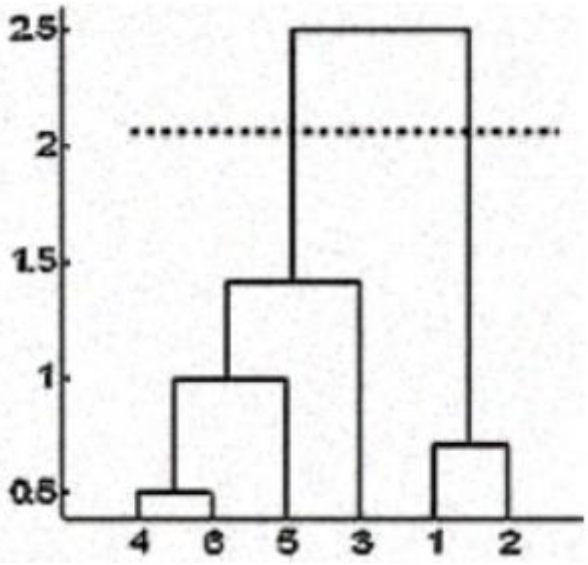


Figure 2.4 Approximation of the number of clusters, using the Horizontal line criterium [11]

The key operation in hierarchical agglomerative clustering is to repeatedly combine the two nearest clusters into a larger cluster. Below is how the method works [15]:

1. It starts by calculating the distance between every pair of observation point and store it in a distance matrix.

2. The method starts uniting the closest pairs of points based on the distances from the distance matrix and as a result the number of clusters goes down by 1.

3. Then it recomputes the distance between the new cluster and the old ones and stores them in a new distance matrix.

4. Lastly it repeats steps 2 and 3 until all the clusters are combined into one single cluster.

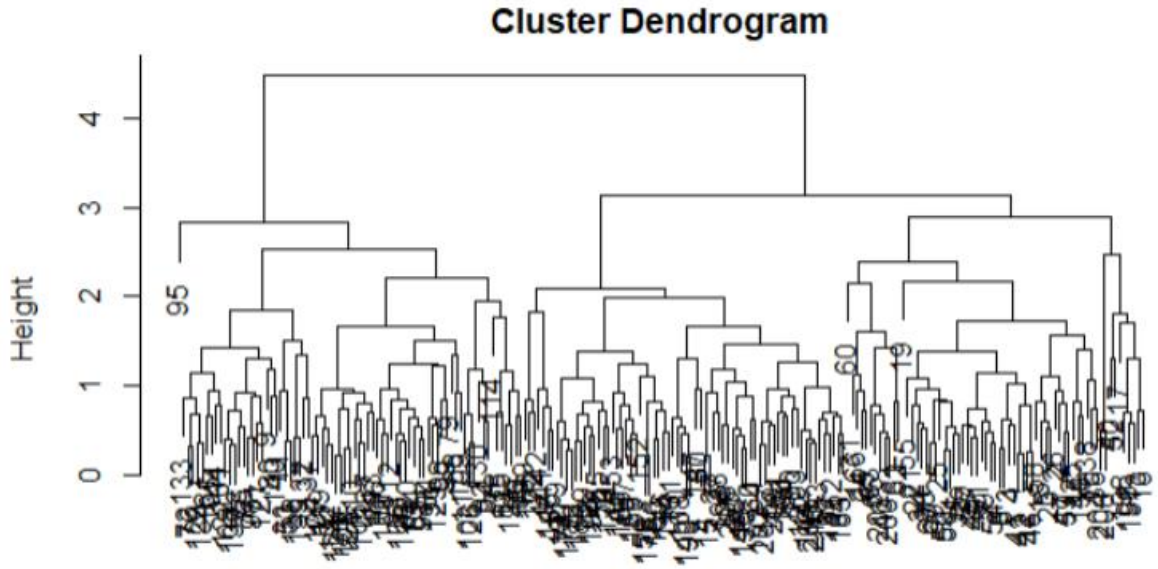


Figure 2.5 Cluster Dendrogram for a big dataset of more than 1000 elements [16].

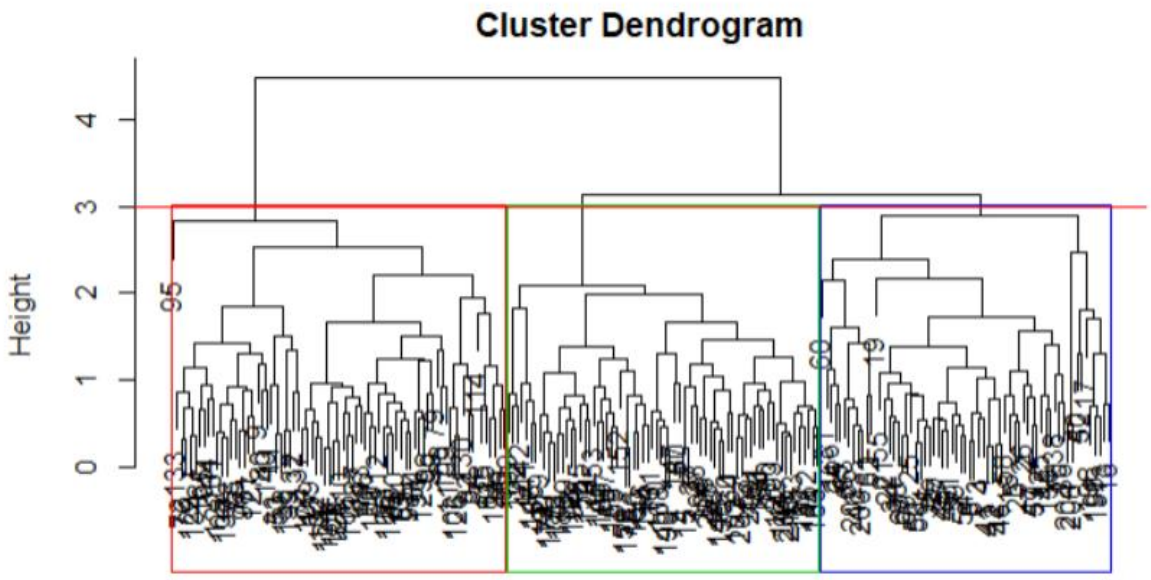


Figure 2.6 Cluster Dendrogram for a big dataset of more than 1000 elements considering 3 clusters represented by different color boxers [16].

Fig. 2.4 and 2.5 show the final cluster dendrogram of a study of more than 1000 elements. As can be seen, the authors of this study decided to use only three clusters, the same ones that are represented by colored boxes in Fig. 2.6

Distance Matrix

The distance matrix should be determined before any analysis is performed. For each step, the matrix is updated to show the distance between each cluster. There are three methods about how the distance between each cluster is measured [17]:

1. Single Linkage: the distance between two clusters is defined as the shortest distance between two points in each cluster.

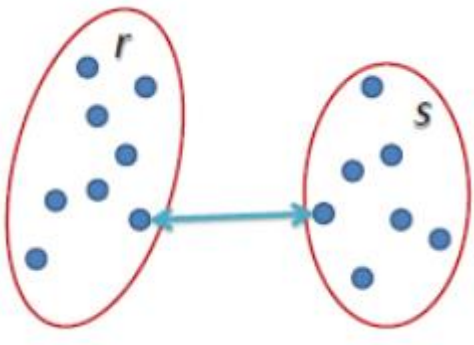


Figure 2.7 Shortest distance between cluster "r" and "s" [20].

$$L(r,s) = \min\big(distance(x_{ri}, x_{sj})\big) \quad (2.2)$$

2. Complete Linkage: the distance between two clusters is defined as the longest distance between two points in each cluster.

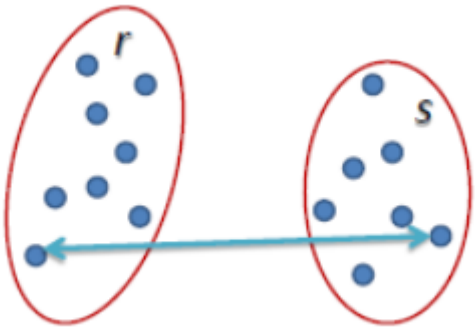


Figure 2.8 Longest distance between cluster "r" and "s" [17].

$$L(r,s) = \max\big(distance(x_{ri}, x_{sj})\big) \quad (2.3)$$

3. Average Linkage: the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.

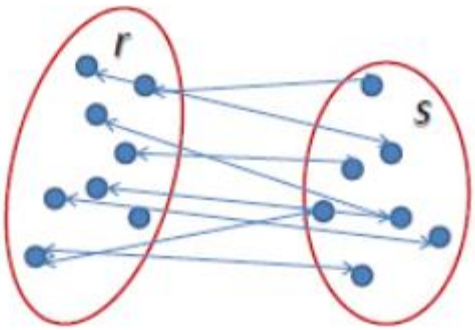


Figure 2.9 Average distance between each point of clusters “r” and “s” [17].

$$L(r,s) = \frac{1}{n_r}\frac{1}{n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} distance\left(x_{ri}, x_{sj}\right) \quad (2.4)$$

DENSITY BASED CLUSTERING METHODS

Density Based Spatial Clustering of Algorithms with Noise (DBSCAN) identifies arbitrarily shaped clusters and outliers (noise) in data. During clustering, DBSCAN identifies points that do not belong to any cluster, which makes this method useful for density-based outlier detection.

The base of the method is that for each point of a cluster the neighborhood with a given radius (Eps) contain at least a minimum number of points (MinPts) [8].

DBSCAN split the data into three classes [8]:

- Core Points: These points are at the interior of a cluster. It means, there are some points in its neighborhood.
- Border Points: There are not some points in its neighborhood, but it falls within the neighborhood of a core point.
- Noise Points: It is a point that is not a core point nor a border point.

To find a cluster, DBSCAN begins with and arbitrary point (p) in a dataset (D) and recovers all the points of the dataset with respect the input parameters Eps and MinPts.

Unlike k-means and k-medoids clustering, DBSCAN does not require prior knowledge of the number of clusters [18].

For certain values of the radius epsilon and the minimum number of points in its neighborhoods, MinPts, the DBSCAN function works like this:

- From the input database D, pick up the first unlabeled observation Xi, and start the first cluster label C to 1.
- Find the set of points the epsilon radius of the actual point. These points are the neighbors.
- Iterate over each neighbor (new points) and repeat the step before until no new neighbors are found in the current cluster C.
- Chose the next unlabeled point in D as the actual point and increase the counter of the cluster by 1.
- For this new point, repeat form step 2 to 4 until all points in D are considered.

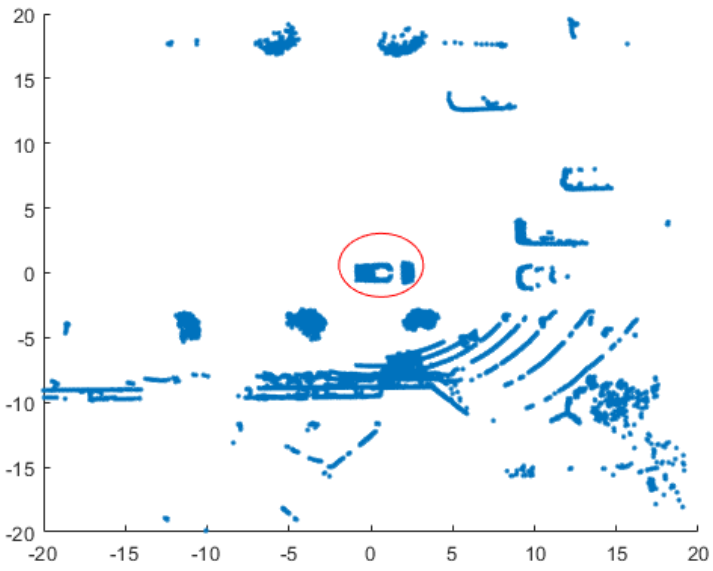


Figure 2.10 DBSCAN method using information about the roof and hood of the vehicles. Look that the center of the set of points are circled in red [18].

Select of the input parameters for DBSCAN method

To select a value for 'MinPts', consider a value greater than or equal to one plus the number of dimensions of the input data [21]. For example, for an n-by-p matrix X, set the value of 'MinPts' greater than or equal to p+1.

One strategy for estimating a value for epsilon is to generate a k-distance graph for the input data X. For each point in X, find the distance to the kth nearest point, and plot sorted points against this distance. The graph contains a knee. The distance that corresponds to the knee is generally a good choice for epsilon, because it is the region where points start tailing off into outlier (noise) territory [21].

K-NEAREST NEIGBORH CLUSTERING

k-nearest neighbor search finds the *k* closest points in your data to a query point or set of query points.

The k-nearest-neighbor classifier is commonly based on the Euclidean distance between a test sample and the specified training samples. [22]

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + \left(x_{ip} - x_{lp}\right)^2} (2.5)$$

Where:

- Xi is an input sample with p features ($x_{i1}, x_{i2,} \ldots, x_{ip}$)
- n be the total number of input samples (i=1,2,…,n) and
- p the total number of features (j=1,2,…,p) .

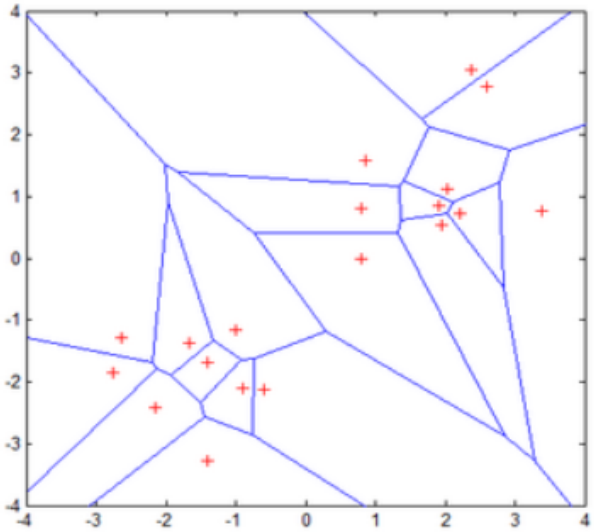


Figure 2.11K-nearest neighbor example: Voroni tessellation showing Voronoi cells of 19 samples marked with a “+” [21].

COMPARISON BETWEEN METHODS

Table 2.1 shows a comparison between the methods considered in this study. The contrast is made based on the mathematical basis of each method, the input parameters, the need to know in advance the number of clusters, and if the method is an outlier detector.

Table 2.1 Comparison table between methods considered [22].

METHOD	BASIS	INPUT PARAMETERS	REQUIRES KNOWING A PRIORI THE NUMBER OF CLUSTERS	Outlier Detector
Hierarchical Clustering	Distance between objects	Pairwise distances between observations	NO	NO
K-Mean Clustering	Distance between objects and centroids	Current observations	YES	NO
Density-Based Spatial Clustering of Algorithms with Noise (DBSCAN)	Density of regions in the data	Current observations or pairwise distances between observations	NO	YES
Nearest Neighbors	Distance between objects	Current observations	NO	DEPENDING ON THE NUMBER OF NEIGHBORS

3 BASIC CONCEPTS OF CLUSTER VALIDATION APPROACHS

3.1 Silhouette value

The silhouette value is a factor used to assess the functioning of a clustering algorithm.

The formula to calculate the silhouette value is the following:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \quad (3.1)$$

Where:

- a (x) is the average distance of sample x (one meter) from all other samples belonging to the same group.
- b (x) is the average distance from sample x to all samples in the closest group [24].

The outcome of s(x) could be from -1 to +1. A positive value close to +1 of s(x) means the sample X is more like other members of its group and X is more different from the members of the other groups. A negative coefficient means the sample X is not like the remaining members of its group, but it is like the members of a different group. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring [25].

The silhouette value can be characterized graphically, using bars representation as the most common representation for its ease of analysis. Each bar represents the Silhouette value of each individual sample. In a good clustering procedure, all the bars must be in the positive side.

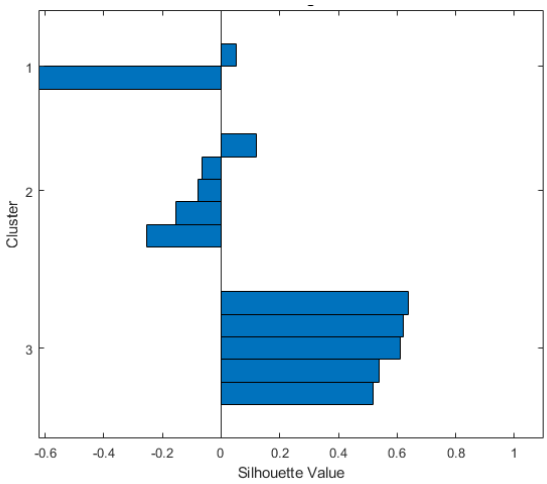


Figure 3.1 Silhouette Value Representation

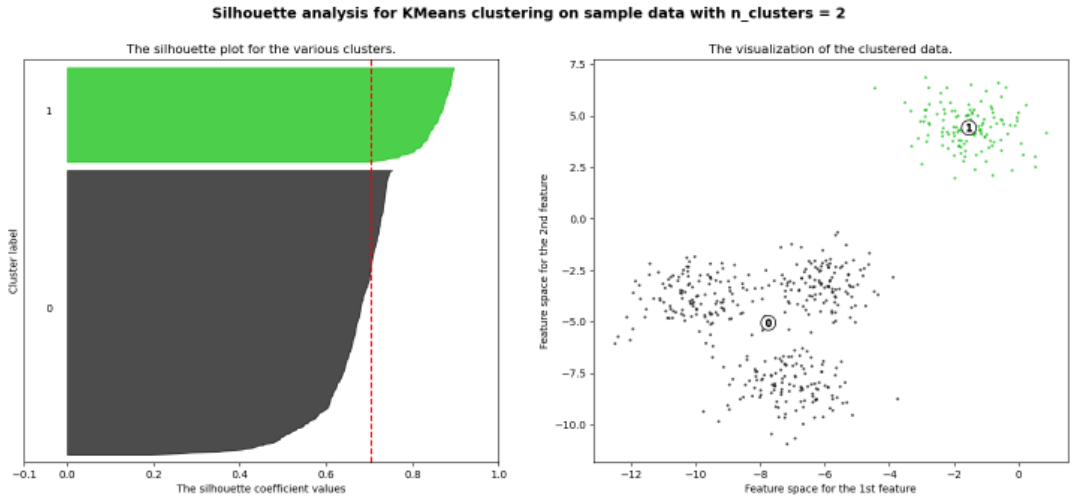


Figure 3.2 Example of Silhouette analysis and dispersion of the cluster data for K means clustering on sample data with k= 2 clusters [25].

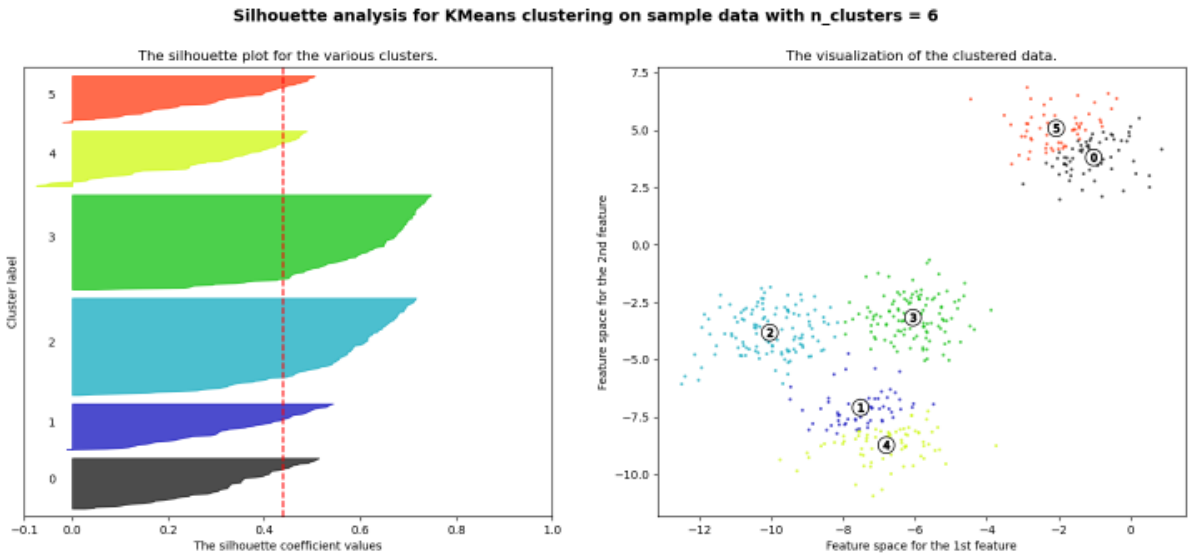


Figure 3.3 Example of Silhouette analysis and dispersion of the cluster data for K means clustering on sample data with k= 6 clusters [25].

SILHOUETTE VALUE IN HIERARCHICAL METHOD USING MATLAB

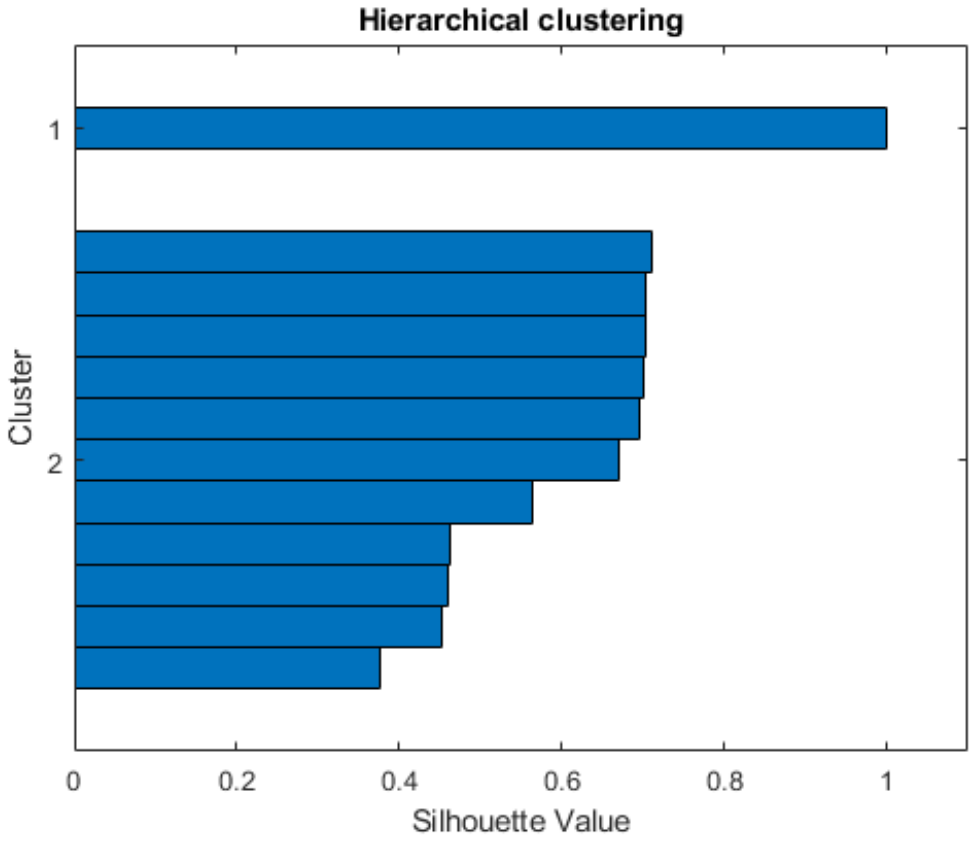


Figure 3.4 Silhouette Value of Hierarchical clustering method, using MATLAB for a set of 12 samples.

For example, the graph above indicates that one of the samples is very different from all the other samples in the dataset, and they all look like each other.

SILHOUETTE VALUE IN DBSCAN METHOD USING MATLAB

In this algorithm the grouping criterion is the epsilon parameter (eps). This parameter is known as the neighborhood distance of a sample. The members of each group must be within the same neighborhood, that is, the distance between each member must be less than epsilon. If a member has no neighbors, it is considered an outlier and a group is created with that sample as the only member. The silhouette plot of this algorithm is:

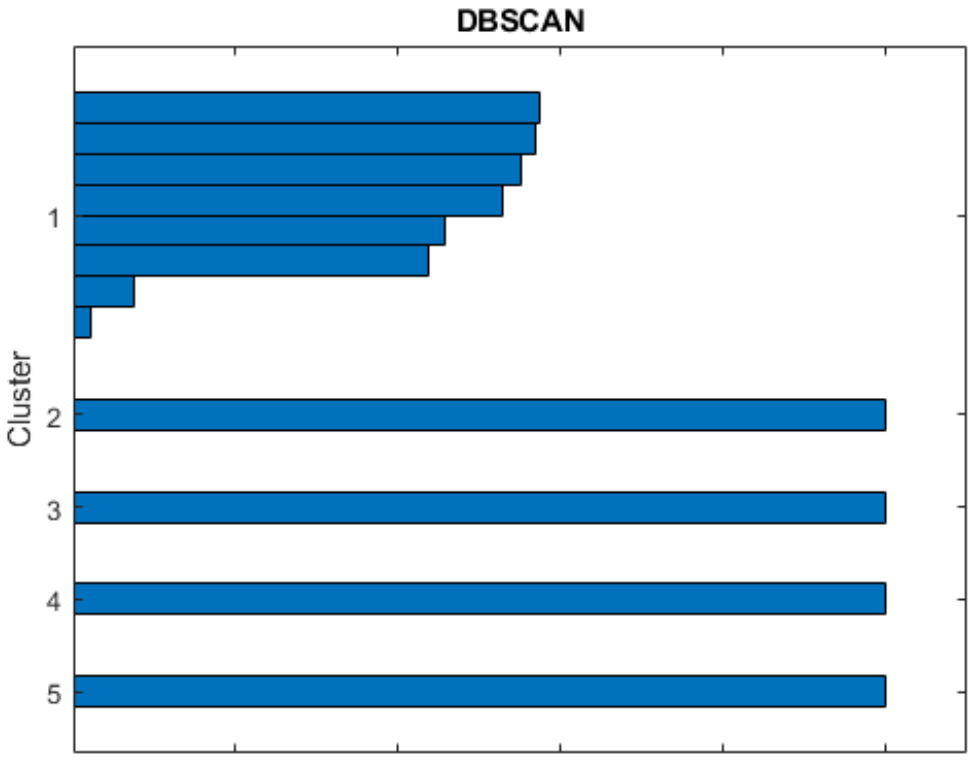


Figure 3.5 Silhouette Value of DBSCAN clustering method, using MATLAB for a set of 12 samples.

For example, the graph above shows that there is a group of 8 members in which all are neighbors or have neighbors in common. Among the members of this group there are two that have little similarity to the other members of the group. The other four groups are from samples that have no neighbors and therefore have only one member. They are identified as outlier by DBSCAN algorithm.

SILHOUETTE VALUE IN K-NEAREST NEIGHBORS METHOD USING MATLAB

For this algorithm, it is needed to have a matrix as a reference, for example a group using K-Mean method of 30 random samples in 6 groups (the *rand_sample* vector). The closest samples (from rand_sample) are compared with the sample to be grouped and assigned to the largest group. The silhouette graph is:

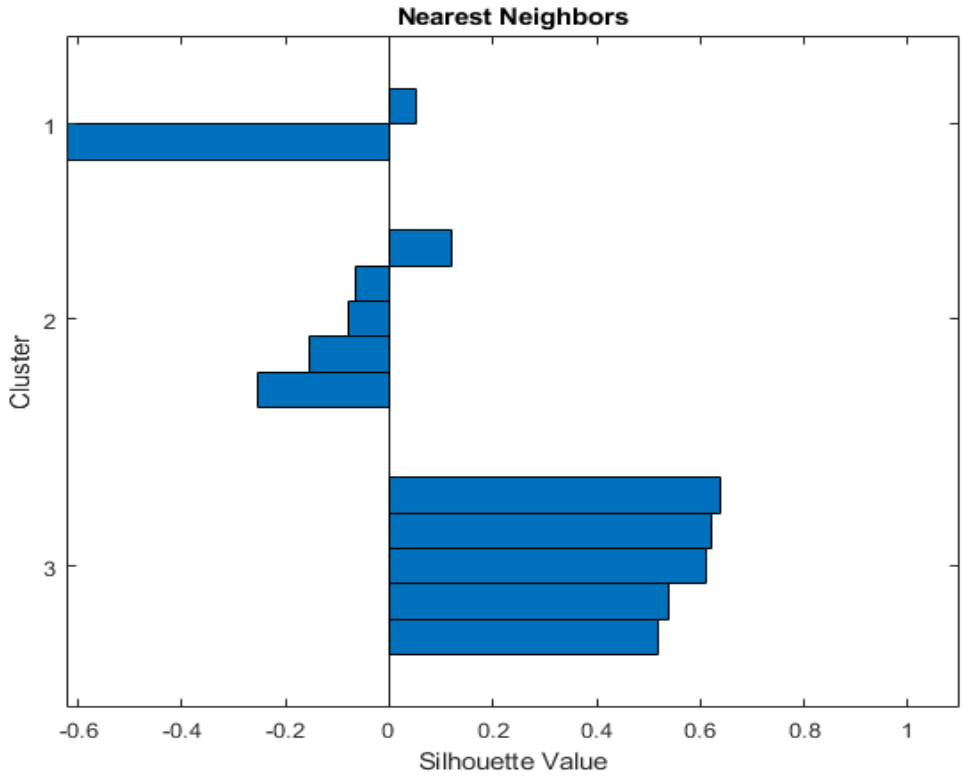


Figure 3.6 Silhouette Value of K-Nearest Neighbors clustering method, using MATLAB for a set of 12 samples.

This graph indicates that 2 of the groups formed do not have members that resemble each other.

SILHOUETTE VALUE IN K-MEAN METHODS USING MATLAB

For this algorithm, the main parameter is the number of clusters K that must be known a priori, having knowledge of the database to be analyzed.

The parameter K can vary from 2 to the size of the sample, but if it is desired to have a good efficiency of the method it is strongly recommended that the number of clusters is not such a high number.

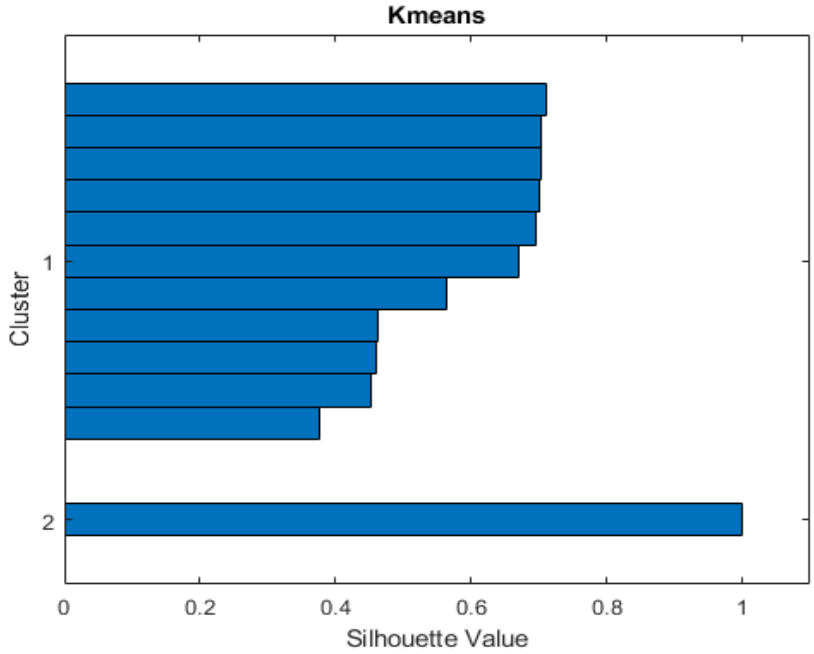


Figure 3.7 Silhouette Value of K-Mean clustering method with k=2 groups, using MATLAB for a set of 12 samples.

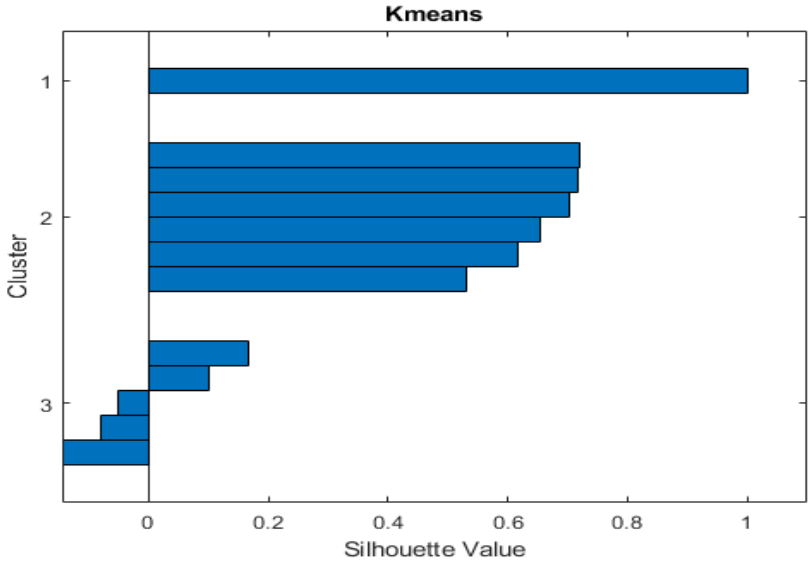


Figure 3.8 Silhouette Value of K-Mean clustering method with k=3 groups, using MATLAB for a set of 12 samples..

As seen in the last two graphs, it is seen that there is a difference between the Silhouette Charts when the number of clusters is changed.

3.2 Average Silhouette value approach

The average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

Average silhouette method calculates the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k [27].

1. The procedure can be calculated as follow:Compute clustering algorithm (for example k-means clustering). A sensitivity analysis (variation of input parameters) can be applied.
2. For each situation, calculate the average silhouette of observations.
3. Plot the silhouette chart according the input parameters.
4. The ubication of the maximum average of silhouette value for each method is considered as the appropriate number of clusters and of the better performance.

The best advantage of using the average silhouette approach. score for finding the best number of clusters is that you use it for un-labelled data set. This is usually the case when running k-means, but it can be applied for other methods [28].

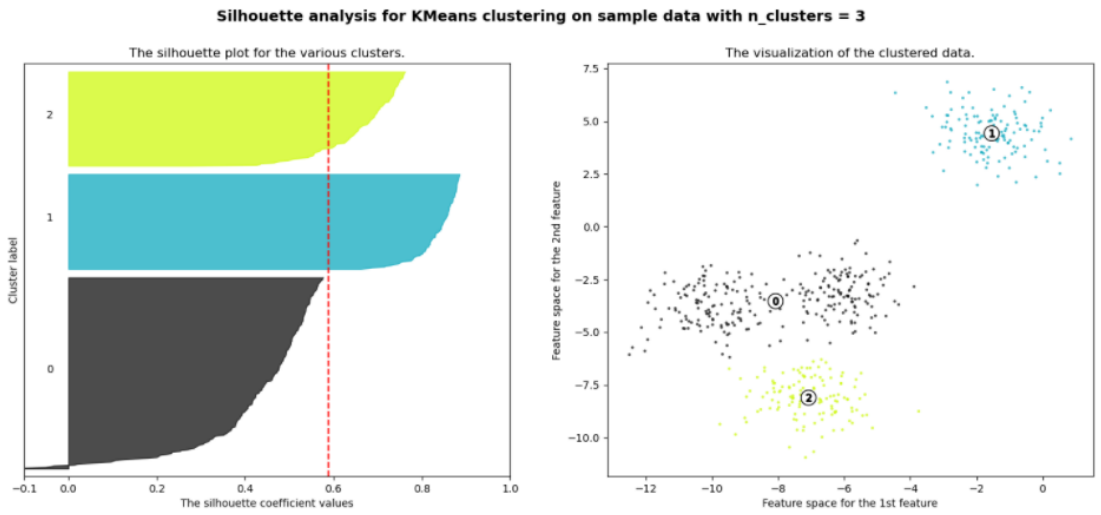


Figure 3.9 Silhouette analysis using K-Mean method with k=3 groups [28].

3.3 Elbow charts approach

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.

There are two kind of procedures to prove the Elbow Charts [32]:

1. Distortion: It is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used.

This method focuses on the percentage of variants as the function of the number of clusters. Based on the idea that there should be an optimal number of k-means algorithm, so adding the number k will not contribute significantly [38]. The value of k is added one by one and the Sum Square Error (SSE) value is recorded.

$$SSE = \sum_{K-1}^{K} \sum_{xi \in Sk} \|X_i - Ck\|_2^2 \quad (3.1)$$

Where SSE is the sum of the average Euclidean Distance of each point against the centroid [39]. When the value drops drastically and forms a smaller angle, then the value of k is found. Starting from k=2 and the SSE value is then added step by step, where kn=k+1, the largest SSE kn - SSE kn − 1 is the point in which the optimal k value is

found. When the value of k is re-added then the new cluster is similar to the previous cluster or the number of errors does not change significantly which resulted in the value of k [33].

2. Inertia: It is the sum of squared distances of samples to their closest cluster center.

To determine the optimal number of clusters, we have to select the value of k at the "elbow" i.e. the point after which the distortion/inertia start decreasing in a linear fashion [32].

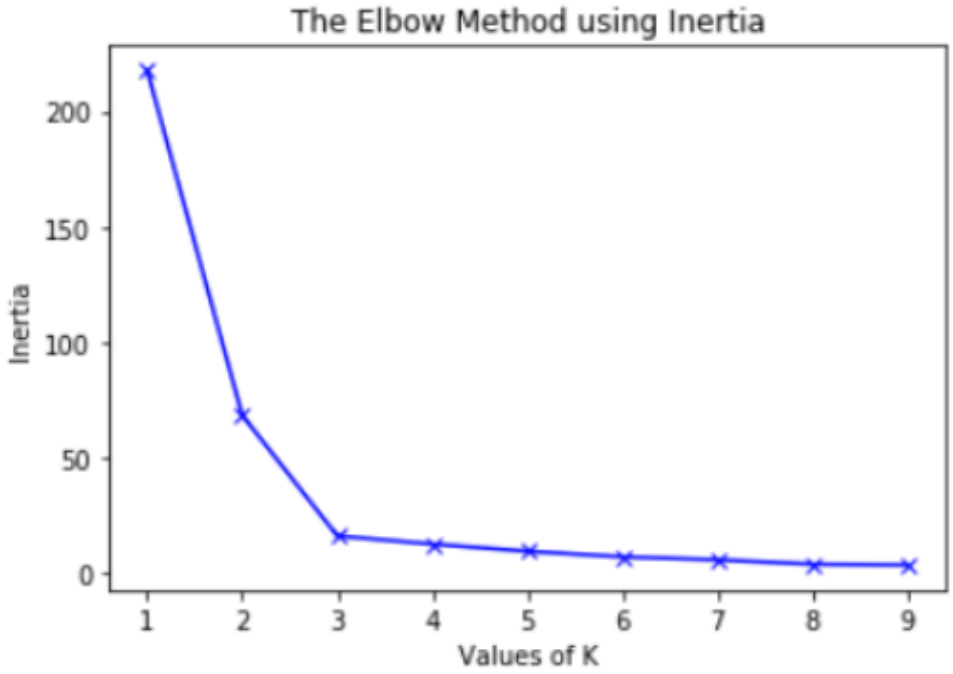


Figure 3.10 The Elbow Chart using the inertia Criteria. There is an Elbow when K= 3 [32].

4 METHODOLOGY OF THE LOAD PROFILE CLUSTERING

4.1 General view of the methodology

The load profile clustering procedures are used to organize customers according to many factors like holiday seasons, temperature, customer infrastructure, etc. as well as the patterns of electricity consumption, and the evaluation of the trend of energy consumed.

The load profile clustering was made using real data of UNARETI, the Electricity and Gas Company of the A2A S.p.A. in Milan. A huge amount data from advanced metering devices was used. The data was measured every 15 minutes during a whole year.

This study utilizes data preparing and dimension reduction methodology (using a filter and PCA), sensitivity analysis based on changing the input parameters of each method, silhouette value analysis and clustering validation. The active load was taken in account in the analysis.

In general, the clustering of the load profiles can mainly be divided into the following steps:

a. Loading raw load data from the database.
b. Load data filter and load data simplification.
c. Analysis of samples and subsequent analysis of the entire database.
d. Sensitivity analysis based on changing the input parameters.
e. Best data representation and best time-step using the samples.
f. Silhouette value analysis and clustering validation, using average silhouette value.
g. Normalization of curves in the selected clustering method

4.2 Loading raw load data from the database

The first step of the study was to collect the information from the measurement devices installed in the UNARETI substations in Milan. More than 1500 meters are installed. The measurement was made every 15 minutes during every day of one year. All the information was stored in a big CVS file. A filter using Microsoft Access was necessary to do.

The information import process using Microsoft Access is as follows:

1. First, a macro to delete the contents of the table was create (the table called “JANUARY_OK” is shown like an example as a part of the whole process) named “DEL”. The process starts by clicking on the macro "DEL".

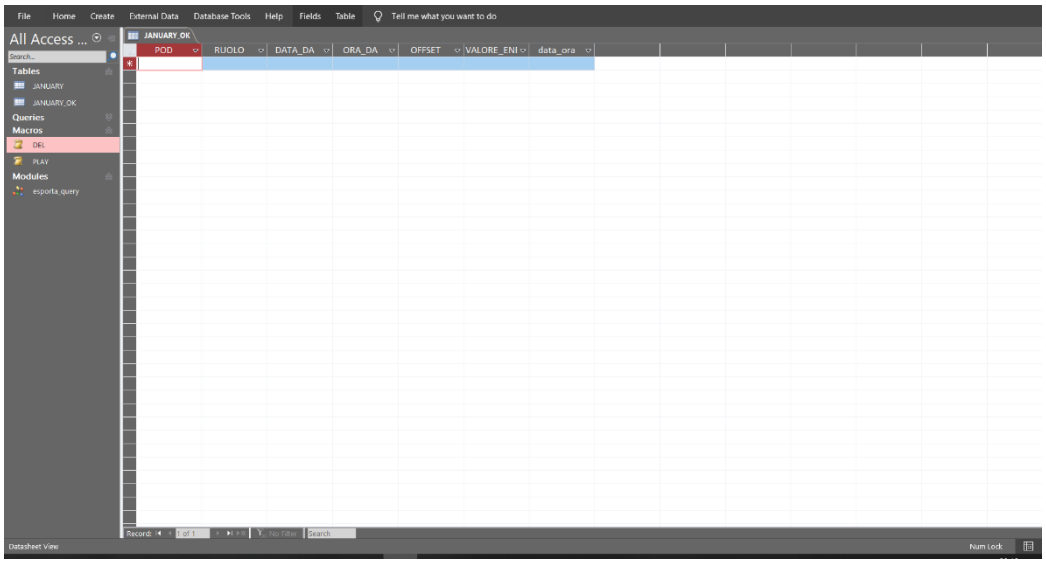


Figure 4.1 Step where the content of the table called “JANUARY_OK” is deleted.

2. Once the table is empty the button “External Data" is chosen from the command ribbon. Then the following sequence is done: “Next Data Source”. “From File”, “Excel"

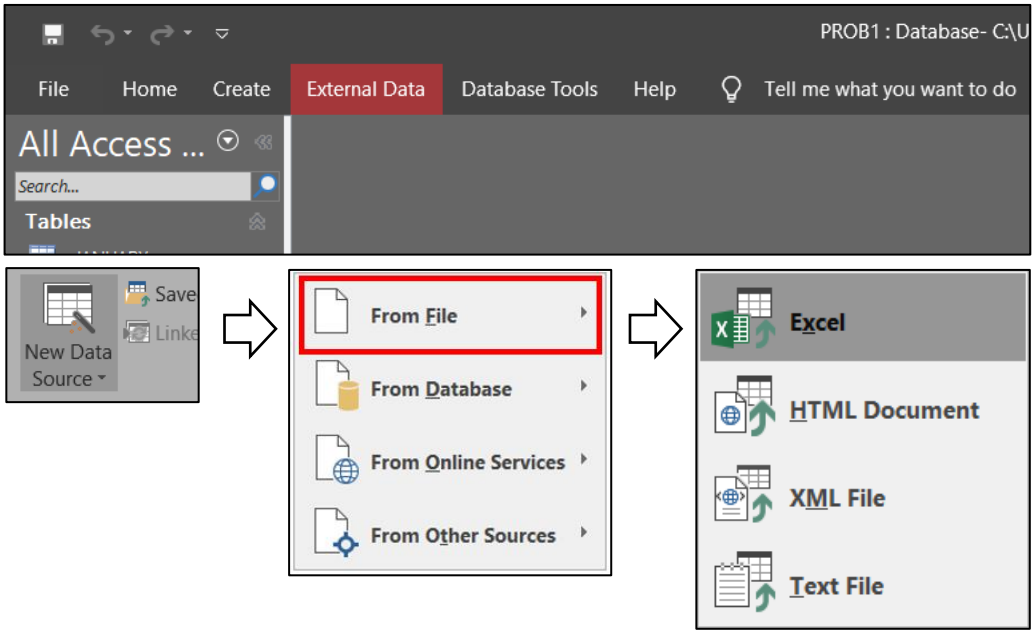


Figure 4.2 Step where the importation of information from the general file is done.

3. Once here, the information is copied into an existing table and select JANUARY_OK

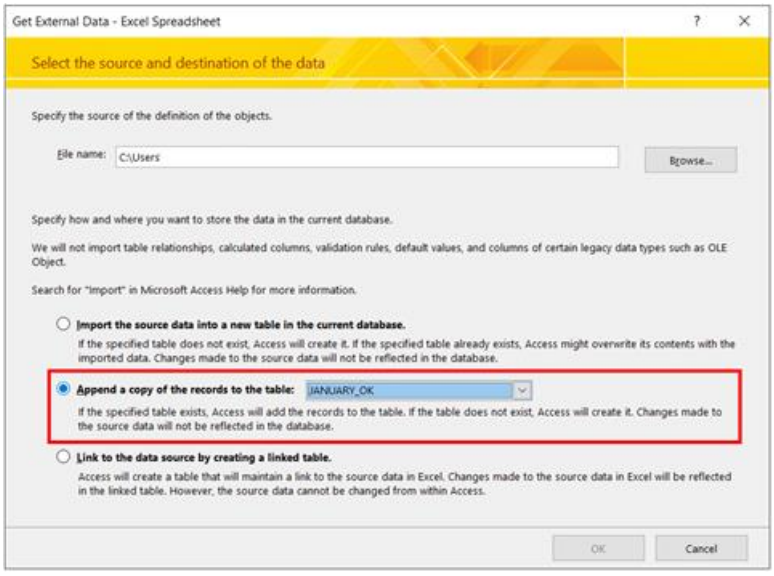
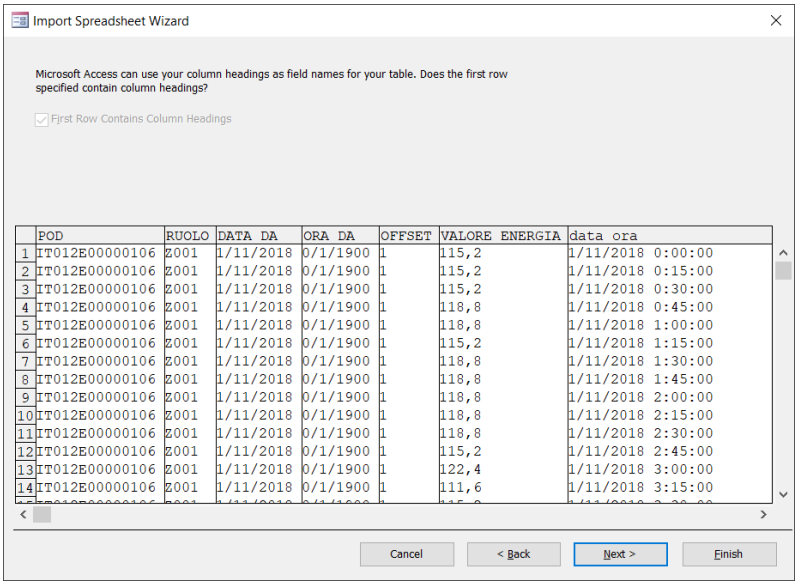


Figure 4.3 Process of copying the information into an existing table.

4. Browse, look for the file of the month and import the file. Once the file is browsed, Then the following sequence is done: “Next”, “Next” and “Finish”.



	POD	RUOLO	DATA_DA	ORA_DA	OFFSET	VALORE_ENERGIA	data_ora
1	IT012E00000106	Z001	1/11/2018	0/1/1900	1	115,2	1/11/2018 0:00:00
2	IT012E00000106	Z001	1/11/2018	0/1/1900	1	115,2	1/11/2018 0:15:00
3	IT012E00000106	Z001	1/11/2018	0/1/1900	1	115,2	1/11/2018 0:30:00
4	IT012E00000106	Z001	1/11/2018	0/1/1900	1	118,8	1/11/2018 0:45:00
5	IT012E00000106	Z001	1/11/2018	0/1/1900	1	118,8	1/11/2018 1:00:00
6	IT012E00000106	Z001	1/11/2018	0/1/1900	1	115,2	1/11/2018 1:15:00
7	IT012E00000106	Z001	1/11/2018	0/1/1900	1	118,8	1/11/2018 1:30:00
8	IT012E00000106	Z001	1/11/2018	0/1/1900	1	118,8	1/11/2018 1:45:00
9	IT012E00000106	Z001	1/11/2018	0/1/1900	1	118,8	1/11/2018 2:00:00
10	IT012E00000106	Z001	1/11/2018	0/1/1900	1	118,8	1/11/2018 2:15:00
11	IT012E00000106	Z001	1/11/2018	0/1/1900	1	118,8	1/11/2018 2:30:00
12	IT012E00000106	Z001	1/11/2018	0/1/1900	1	115,2	1/11/2018 2:45:00
13	IT012E00000106	Z001	1/11/2018	0/1/1900	1	122,4	1/11/2018 3:00:00
14	IT012E00000106	Z001	1/11/2018	0/1/1900	1	111,6	1/11/2018 3:15:00

Figure 4.4 Information in the browsing stage.

5. Once the process is finished, the table in Microsoft Access contains the data.

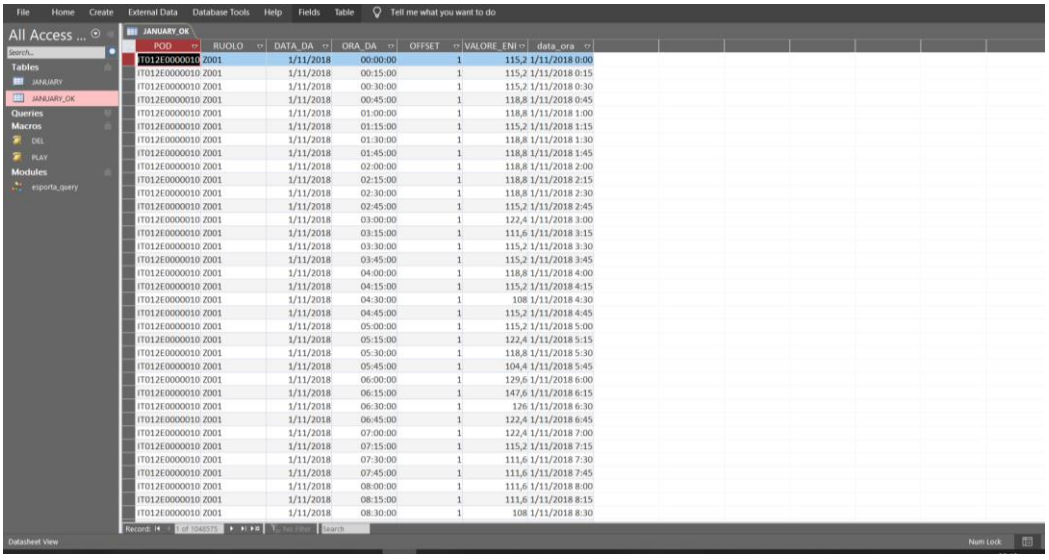


Figure 4.5 Table with the information processed using Microsoft Access.

6. For each month file, perform the steps from the second to the fifth.
7. At the end the entire database, for the 12 months of the year and for the respective number of days of each month. The information was stored in Excel files. This step was made by using a MATLAB code.

4.3 Load data filter and load data simplification

Once the 365 excel files have the entire database distributed, a Principal Component analysis was applied in order to simplify the dimension the database. The following code line in MATLAB was used in order to run the PCA and filter the database [34]:

$$[coeff, score, latent, tsquared, explained, muM] = pca(\dots) \quad (4.1)$$

$$matrix(find(sum(isnan(matrix), 2), :) = [\,] \quad (4.2)$$

Because of possible missing values in the entire dataset with this code line, the principal component can be performed using the ALS (alternating least squares) algorithm.

Equation 4.1 returns the following:

- The principal components of a database in a matrix representation. They are saved in *coeff variable.*
- Principal component scores. They are the representations of a matrix in the principal component space. Rows of score correspond to observations, and columns correspond to components. They are saved in score variable.
- The principal component variances are saved in latent variable.
- The percentage of the total variance explained by each principal component, returned as a column vector *("explained")*
- The percentage of the total variance explained by each principal component, *Estimated* means of the variables in X, returned as a row vector when Centered is set to true.

When Centered is false, the software does not compute the means and returns a vector of zeros (“mu”), the estimated mean of each variable in X.

4.4 Analysis of samples and subsequent analysis of the entire database

Before working with the entire database, an analysis was done with 3 different sets of samples. Each set of samples was different from each other. Each set contained 20 measurements from different UNARETI substations. The measurements were randomly selected, using MATLAB codes.

The results obtained with the samples will be replicated to the entire dataset.

4.5 Sensitivity analysis based on changing the input parameters

The sensitivity analysis was performed first for the set of samples and then with this experience, it was applied to the entire database. On the side of the sets of samples, the aim is to know which combination of the input parameters is the best for maximizing the average Silhouette value of each method.

The sensitivity analysis is based on the input parameter for each clustering method, and how they can affect the final average Silhouette value for each cluster. For this analysis, those single member groups with an average silhouette value equal to one, are not considered.

A MATLAB code was developed to change each input parameter, and automatically at the end of each analysis obtain which combination has the highest silhouette value.

In the study, four methods are analyzed (DBSCAN, Hierarchical, K Nearest Neighbor, K-Mean). Each method is shown below with its respective input parameters to be changed, in order to maximize its silhouette value:

1. DBSCAN: In this method, there are two input parameters to be changed:
   - Variable epsilon, known like “ϵ”. It is the neighborhood distance of a sample.
   - The variable Minimum of Points (variable “MinPoints” through the study). It is the minimum number of samples within a specific neighborhood.

2. Hierarchical, the input parameter to be changed is the buffer size (variable “d” through the study).
3. K-Nearest Neighbor uses as input parameters the number K of the closest neighbors to each point (every measurement for each time step.
4. K-Mean (and its variants), the input data to be considered is the number of clusters K.

The above considerations must be taken both for the different sets of samples, and for the entire database.

4.6 Best data representation and best time-step using the samples

a. Best data representation

Once the best pattern of input parameters had been selected using the samples information, the analysis to choose the best data representation should be performed. Three data representations are considered: absolute values, and p.u. values using the maximum power of the entire set as a reference value, and other p.u. values but using as reference value the maximum power of each individual sample.

The process of getting the best representation between the three representation is to get the average silhouette value without considering those single member groups with silhouette value equal to 1. Then, for different combinations of each method (DBSCAN, HIERARCHICAL, K NEAREST NEIGHBOR, K MEAN-K=2, K MEAN -K=3, K MEAN -K= 4, K MEAN -K= 5, etc.) and each time step (15 minutes, hourly and daily time step) some bar charts should be made.

This process should be done each time when different sets of samples are used, where its members must be chosen at random. The conclusions obtained in this process will be replicated to the entire database.

b. Best time step

Once the best pattern of input parameters had been selected, a similar process as selecting the best data representation should be taken, it will be an analysis to choose the best time-step. Three time-step representation are considered through the study: 15 minutes time step, hourly and daily representation. The last two representations were obtained considering the mean of the 15-minute data in the corresponding time.

The process of getting the best representation between the three time-step (15 min, hourly and daily time step) was to get the average silhouette value without considering those single member groups with silhouette value equal to 1. Then, for different combination of each method (DBSCAN, HIERARCHICAL, K NEAREST NEIGHBOR, K MEAN-K=2, K MEAN -K=3, K MEAN -K= 4, K MEAN -K= 5, etc.) and each representation (absolute values, and the two P.U. values) some bar charts should be made.

Similar to best data representation analysis, this process should be done each time when different sets of samples are used, where its members must be chosen at random. The conclusions obtained in this process will be replicated to the entire database.

4.7 Silhouette value analysis and cluster validation

Once the best combination of the data representation and time step is obtained, thanks to the analysis of the sample sets, and their subsequent application to the entire database; the next step is to search among all the methods which one offers the best performance, based on the silhouette value. At this time, a sensitivity analysis should also be performed, to confirm the behavior shown in the analysis of the samples and to select the best arrangement of input parameters to get the maximum silhouette value. A MATLAB code will be used.

Remember that the silhouette value is a factor used to assess the functioning of a clustering algorithm. Its outcome is a value between -1 and +1. A positive value close to +1 of s(x) means the sample X is more like other members of its group and X is more different from the members of the other groups. A negative coefficient means the sample X is not like the remaining members of its group. The most adequate and used representation of the silhouette value is by using bar charts.

The basis of decision to know which is the best clustering method among those considered, is to calculate the validation clustering criterium, it should be average silhouette value. The average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

4.8 Normalization of curves for the clustering methods

Once the best representation, the best time-step representation, and the best method are founded using the previous analysis and the most important part of the study is to get the normalized load curves of each curve.

To obtain this curve, the per unit representation will be used. The reference value would be the maximum power of the cluster.

This is done because for planning purposes, this curve is important to have knowledge of the most representative curve of each group.

5 RESULTS

5.1 PCA analysis

In this section, most of the data from our studies are represented mainly by the first two PCA components. The 1st component has 88.44% and 2nd component has 2.68% of the information. This confirms the way in which the methods work, i.e., the tendency to create a large group of many members. This information is gotten applying the following MATLAB code over the data.

$$[coeff, score, latent, tsquared, explained, mu] = pca(matrix) \quad (5.1)$$

According to MathWorks help center, explained from equation (5.1) returns the percentage of the total variance explained by each principal component [29].

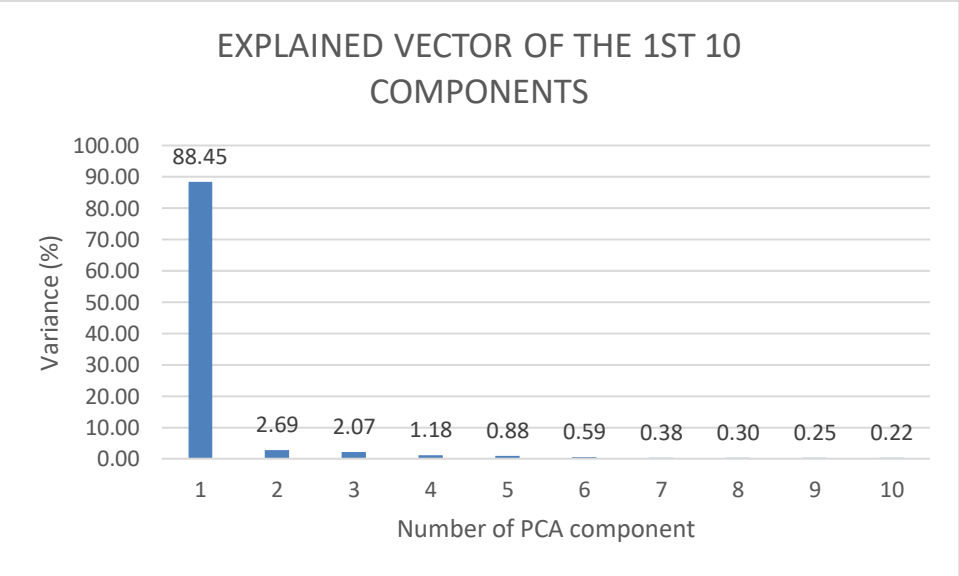


Figure 5.1 Result of applying explained vector from PCA function. The 10 first PCA components are shown.

5.2 Sensitivity Analysis of the Samples

A sensitivity analysis of the three randomly sets of 20 samples were applied in order to know which are the best combinations of the input parameters for maximizing the average Silhouette value of each method.

The sensitivity analysis is based on the input parameter for each clustering method, and how they can affect the final average Silhouette value for each cluster. For this analysis, those single member groups with an average silhouette value equal to one, are not considered.

Four methods are used for this analysis: DBSCAN, Hierarchical, K Nearest Neighbor and K Mean (and its variants). In the case of DBSCAN method, there are two input parameters to be changed, ϵ (variable "epsilon" in the next tables and charts) and the Minimum of Points (variable "MinPoints" in the next tables and charts). On the other hand, the Hierarchical method, the input parameter to be changed is the buffer size (variable "d" in the next tables and charts). K-Nearest Neighbor uses as input parameters the number K of the closest neighbors to each point (every measurement for each time step). Finally, the K Mean method, the input data to be considered is the number of clusters K.

Additionally, a MATLAB code was made in order to select the adequate pattern of input parameters for maximizing the average silhouette value. This code was run for each of the three sets of samples to have a general idea which input parameter fit in the sample analysis and which could be used for the whole data.

The best arrangement of input parameters for each method and its average silhouette value are gotten for different data representation (absolute values, and two P.U. values) and for different time-step (15 minutes, hourly and daily time step). Finally based in the previous selection, the best approach for each of the three set of samples is selected.

5.3 Best value representation

Once the best pattern of input parameters had been selected, the analysis to choose the best value representation was performed. Based on the results about the average Silhouette value of three different sets of 20 samples considering the three representations (absolute values, and the two P.U. values), it can be said that the best results gotten is when the absolute values are used. The process of getting the best value representation between the three representation was to get the average silhouette value without considering those single member groups with silhouette value equal to 1. Then, for different combinations of each method (DBSCAN, HIERARCHICAL, K NEAREST NEIGHBOR, K MEAN-K=2, , K MEAN -K=3, K MEAN -K= 4, K MEAN -K= 5) and each time step (15 minutes, hourly and daily time step) some bar charts were made(they are shown below). It means 21 charts were evaluated. In a single chart, the behavior of the three set of samples and the three-value representation were located (9 bars in one chart).

An additional consideration was taken: the analysis for number of clusters bigger than 5 were not considered because its average Silhouette value were few lower that the other cases.

The charts and analysis about the results of the three set of samples is shown as following:

DBSCAN METHOD CHARTS

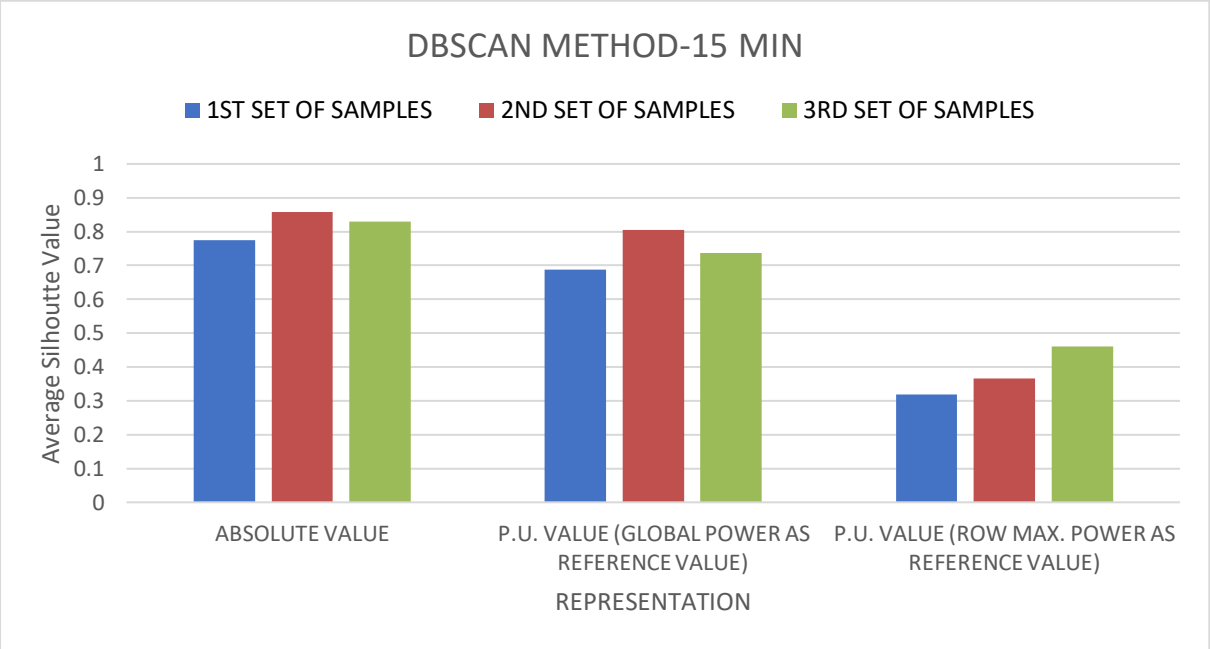


Figure 5.2 Chart of the Average Silhouette value vs the data representation. DBSCAN method and 15 minutes time step for the three different set of samples are considered.

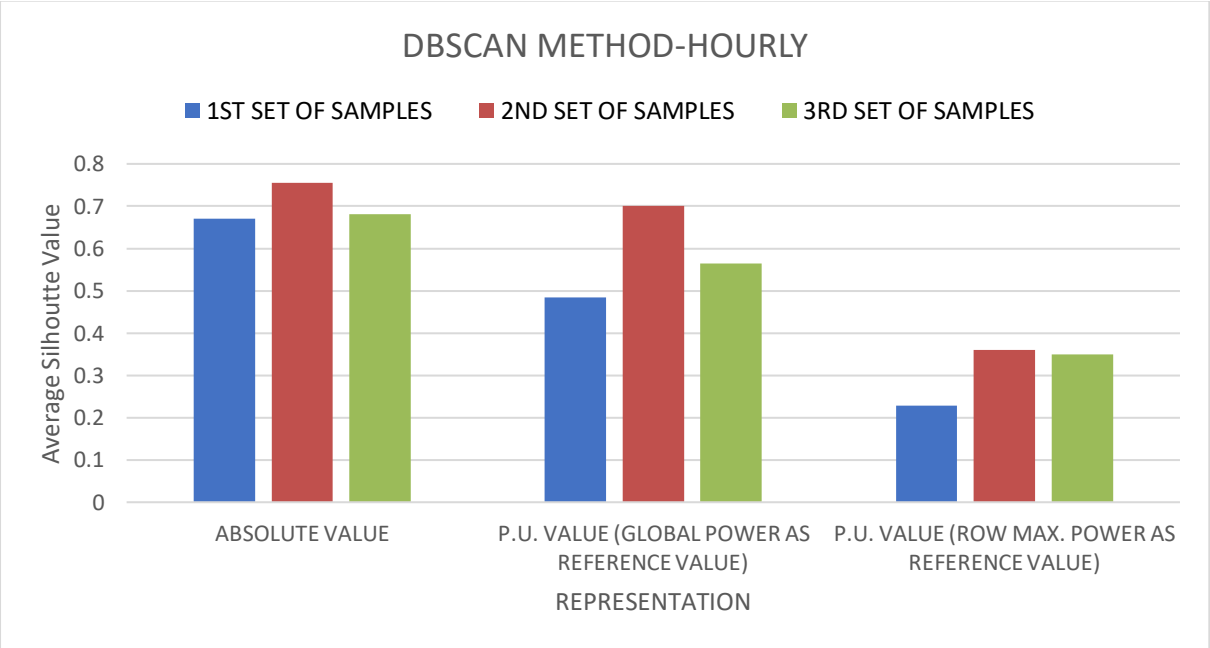


Figure 5.3 Chart of the Average Silhouette value vs the data representation. DBSCAN method and Hourly time step for the three different set of samples are considered.

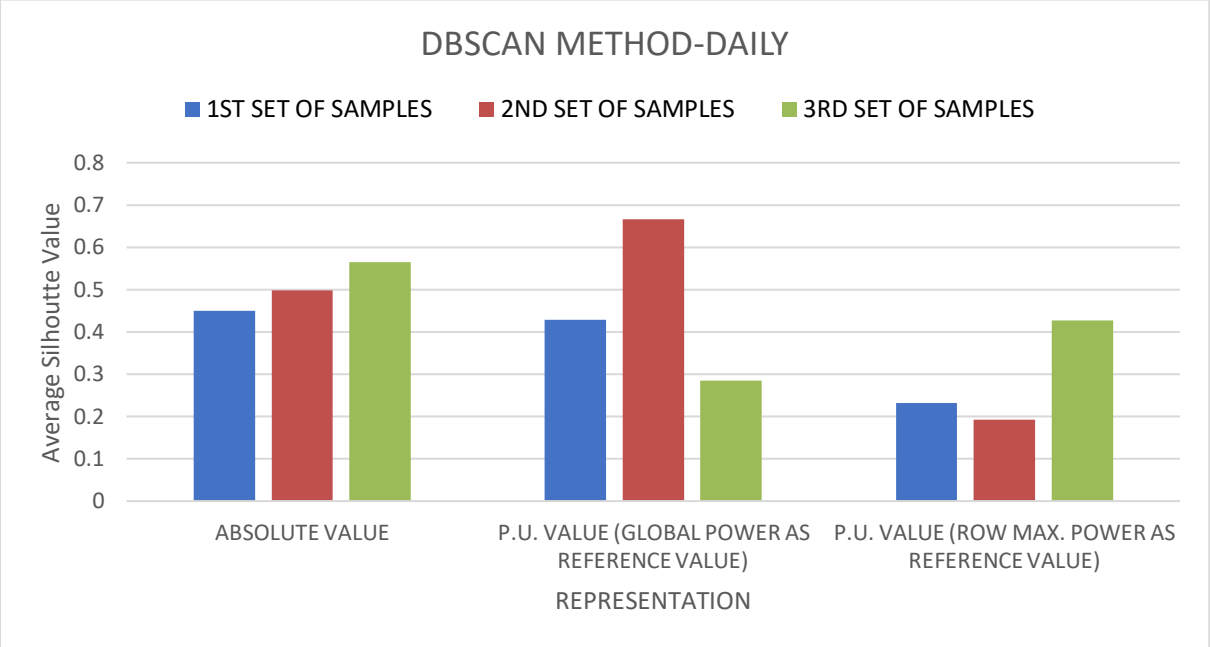


Figure 5.4 Chart of the Average Silhouette value vs the data representation. DBSCAN method and Daily time step for the three different set of samples are considered.

HIERARCHICAL METHOD CHARTS

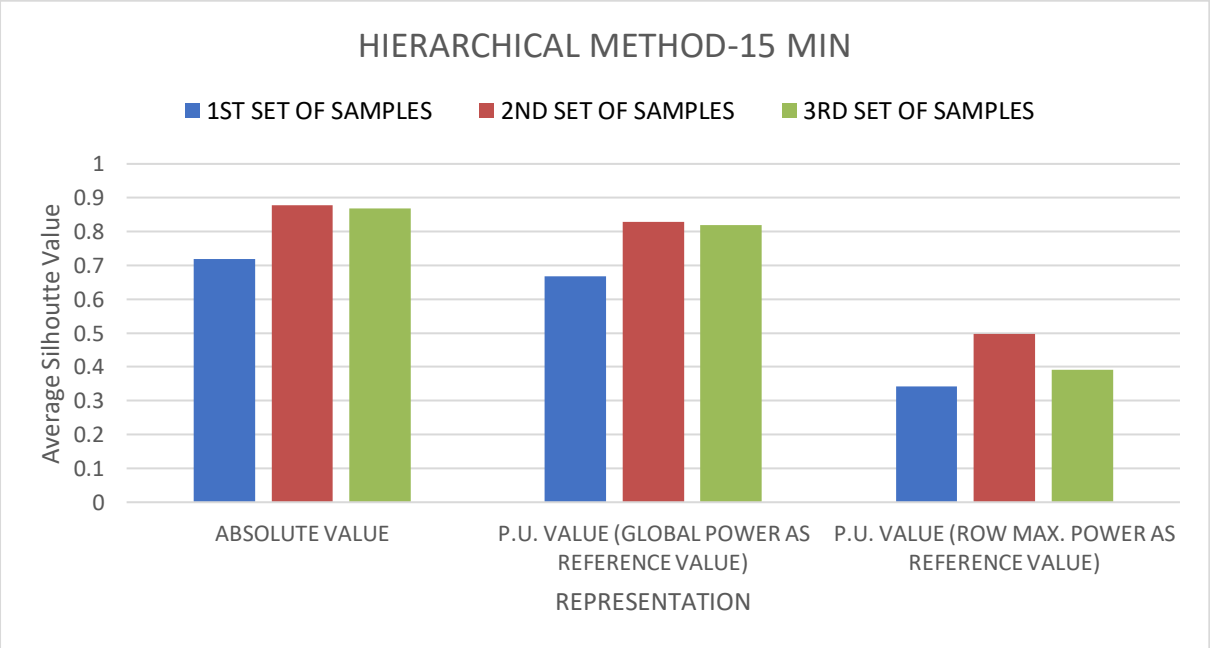


Figure 5.5 Chart of the Average Silhouette value vs the data representation. HIERARCHICAL method and 15 minutes time step for the three different set of samples are considered.

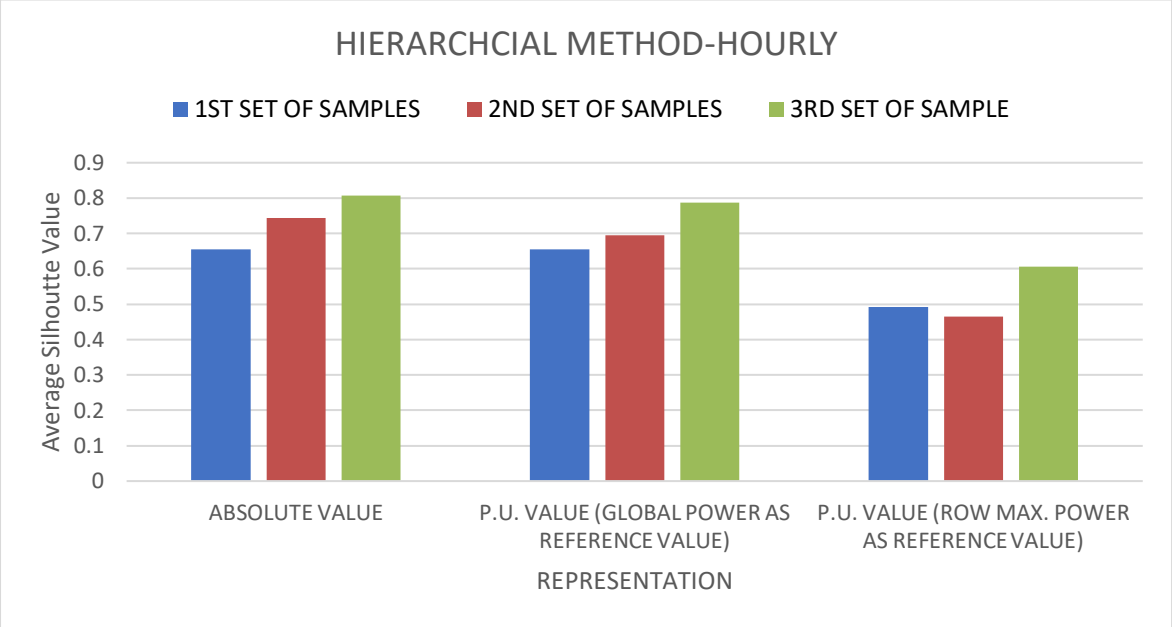


Figure 5.6 Chart of the Average Silhouette value vs the data representation. HIERARCHICAL method and Hourly time step for the three different set of samples are considered.

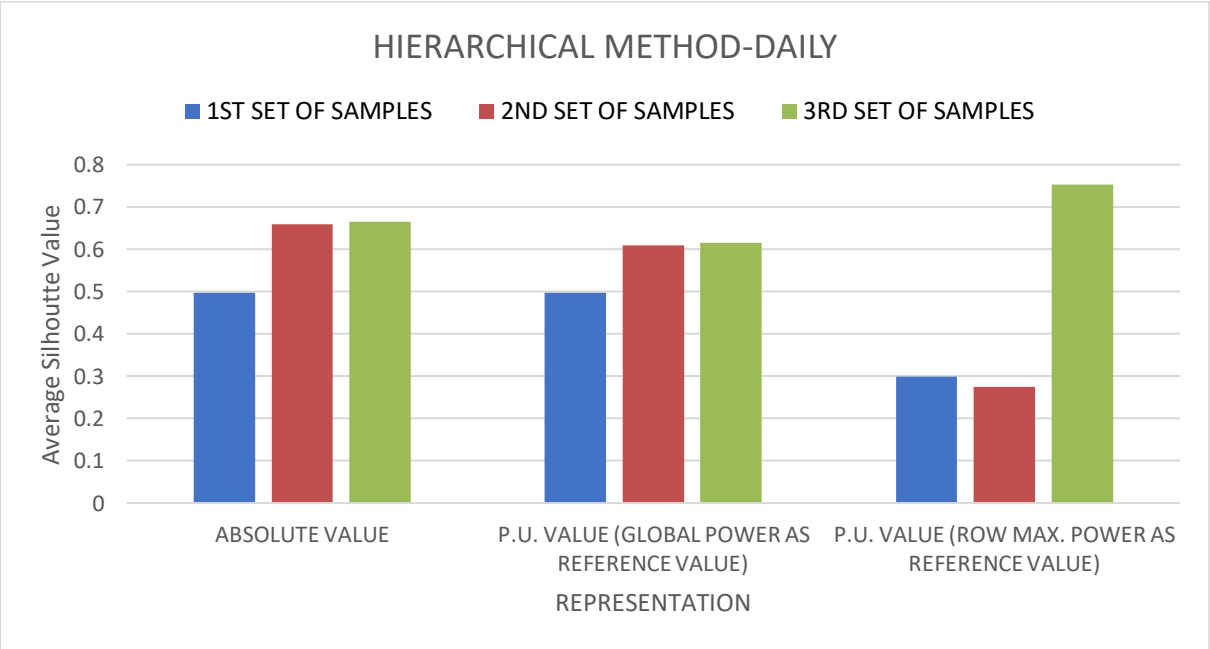


Figure 5.7 Chart of the Average Silhouette value vs the data representation. HIERARCHICAL method and Daily time step for the three different set of samples are considered.

K-NEAREST NEIGHBOR METHOD CHARTS

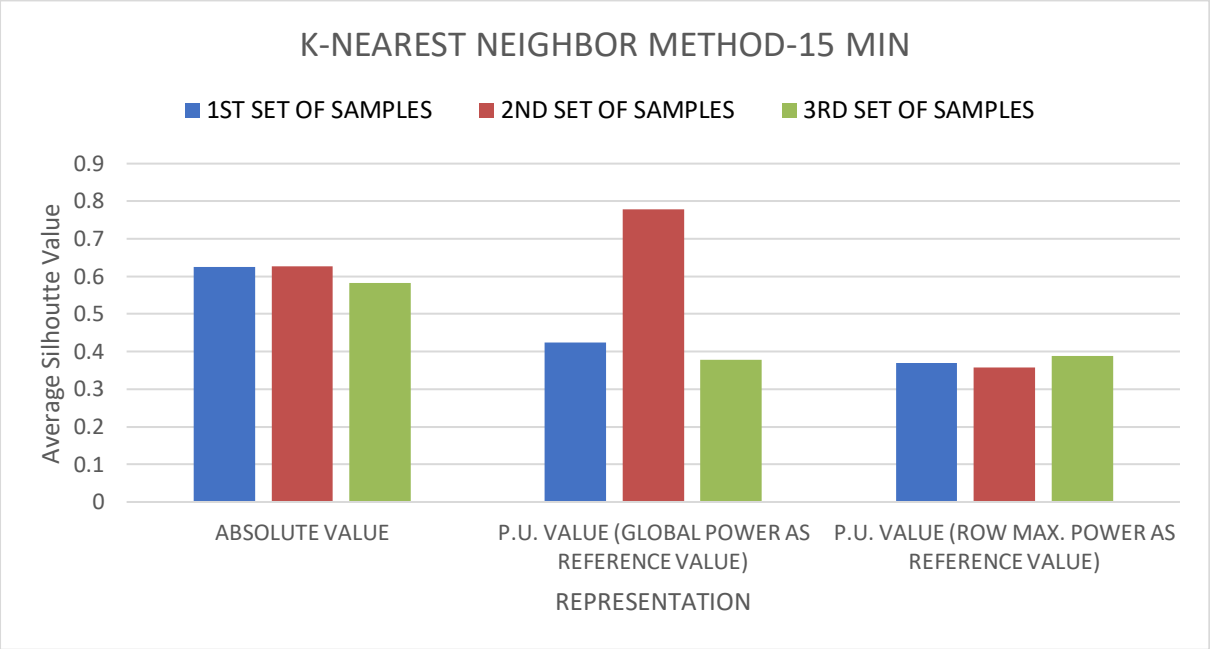


Figure 5.8 Chart of the Average Silhouette value vs the data representation. K-Nearest Neighbor method and 15 minutes time step for the three different set of samples are considered.

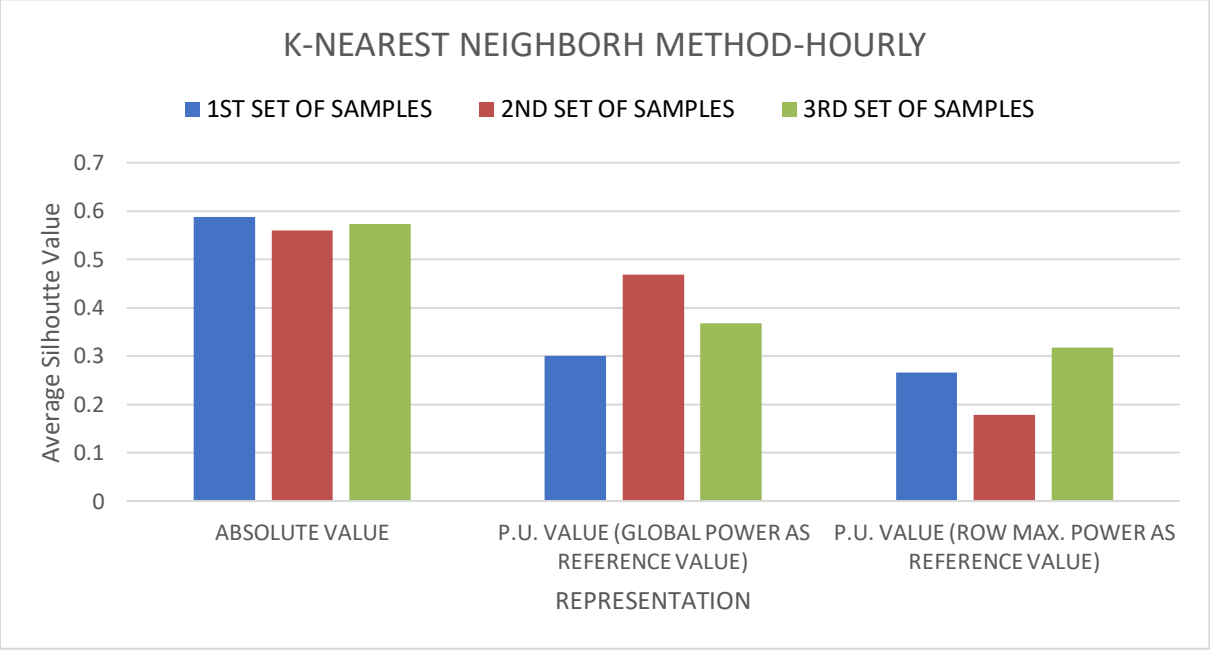


Figure 5.9 Chart of the Average Silhouette value vs the data representation. K-Nearest Neighbor method and Hourly time step for the three different set of samples are considered.

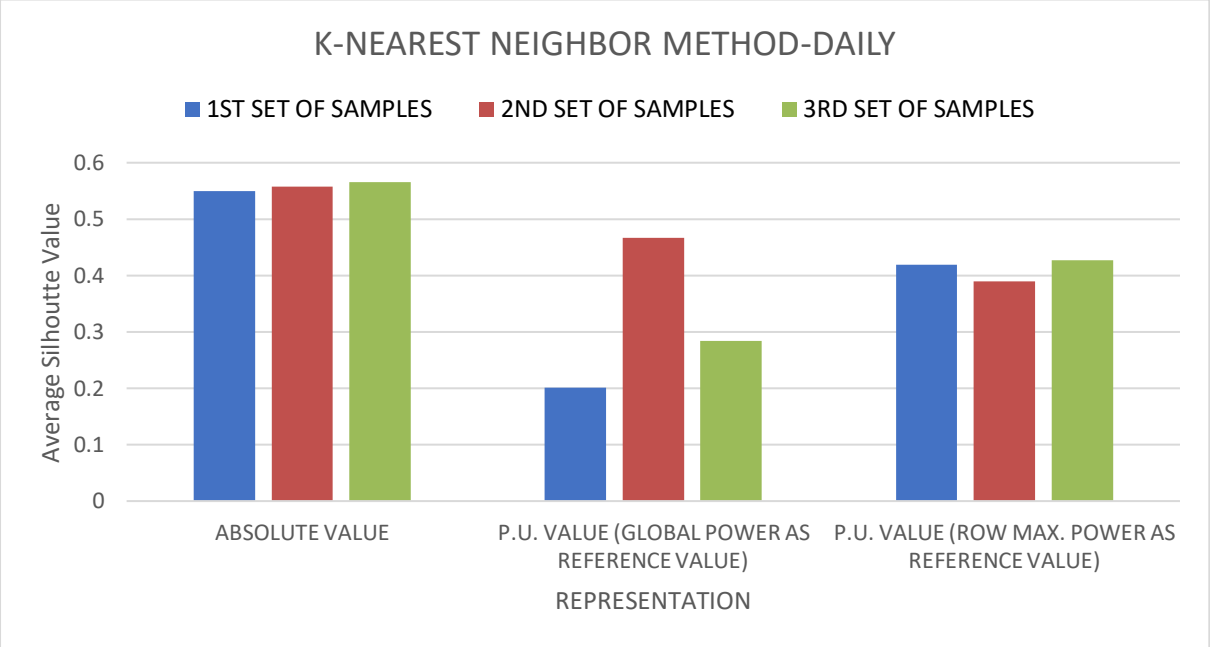


Figure 5.10 Chart of the Average Silhouette value vs the data representation. K-Nearest Neighbor method and Daily time step for the three different set of samples are considered.

K-MEAN (K=2 CLUSTERS) METHOD CHARTS

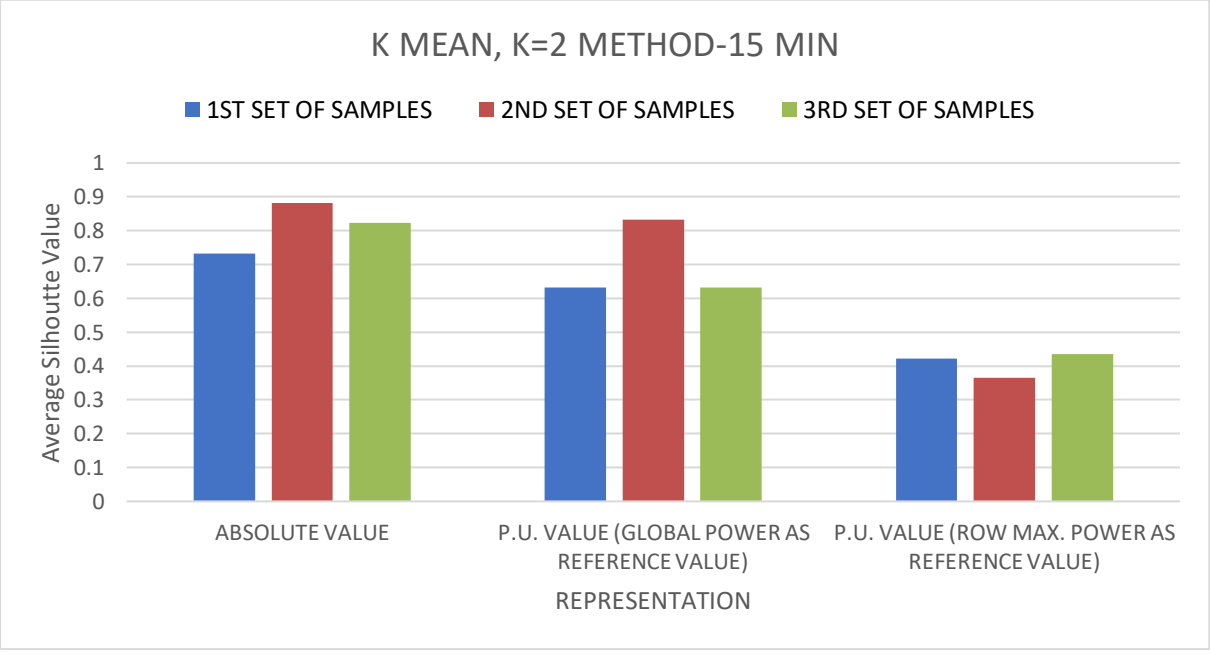


Figure 5.11 Chart of the Average Silhouette value vs the data representation. K-Mean (K=2) method and 15 minutes time step for the three different set of samples are considered.

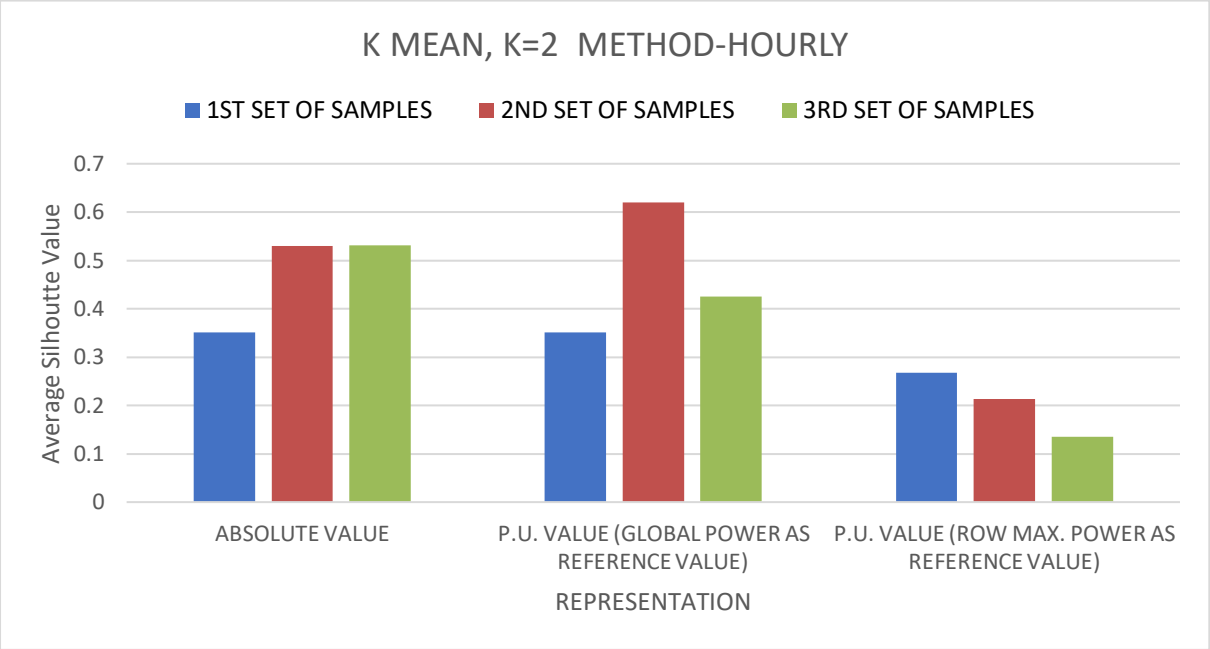


Figure 5.12 Chart of the Average Silhouette value vs the data representation. K-Mean (K=2) method and Hourly time step for the three different set of samples are considered.

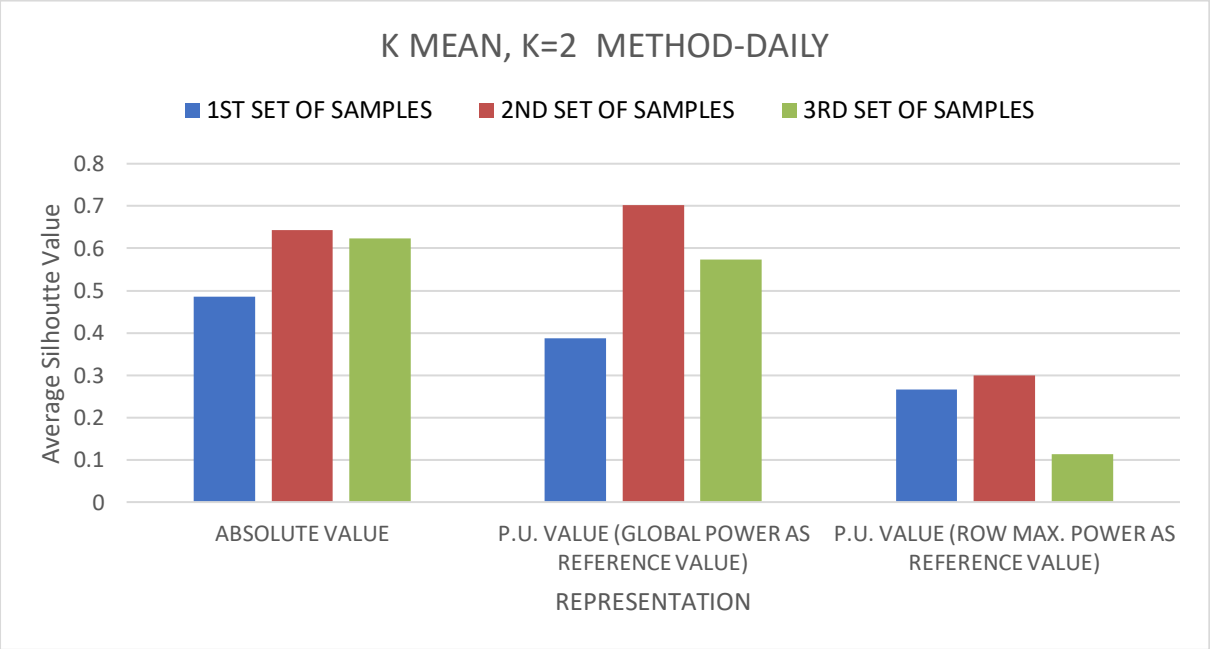


Figure 5.13 Chart of the Average Silhouette value vs the data representation. K-Mean (K=2) method and Daily time step for the three different set of samples are considered.

K-MEAN (K=3 CLUSTERS) METHOD CHARTS

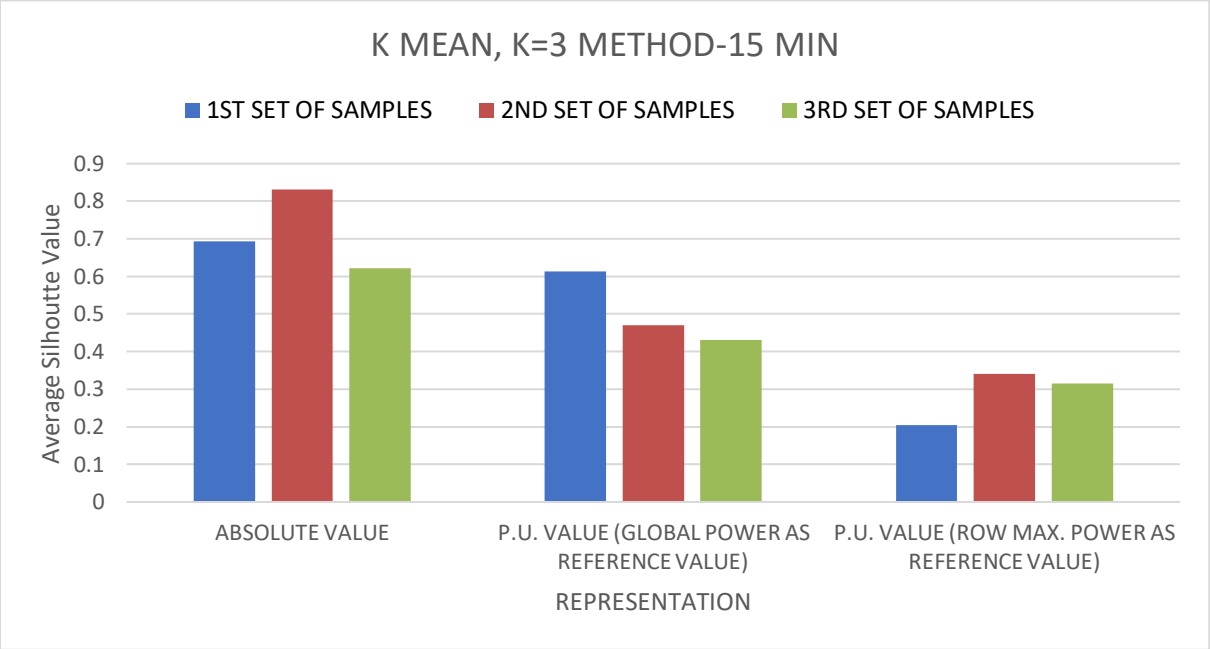


Figure 5.14 Chart of the Average Silhouette value vs the data representation. K-Mean (K=3) method and 15 minutes time step for the three different set of samples are considered.

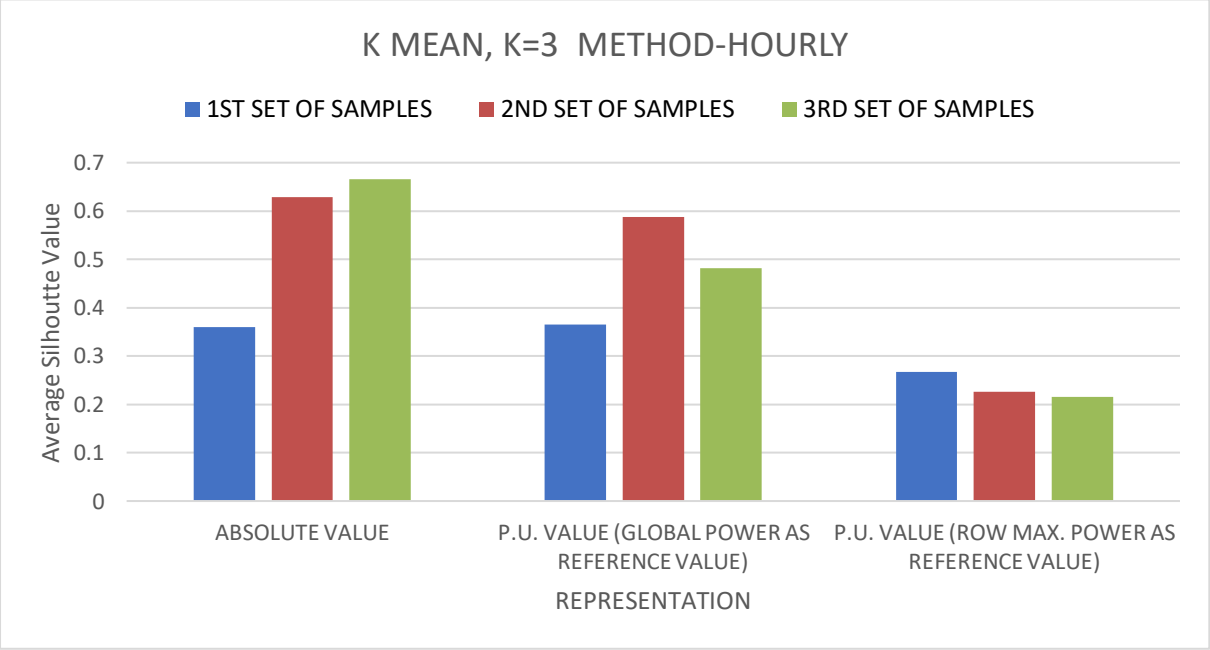


Figure 5.15 Chart of the Average Silhouette value vs the data representation. K-Mean (K=3) method and Hourly time step for the three different set of samples are considered

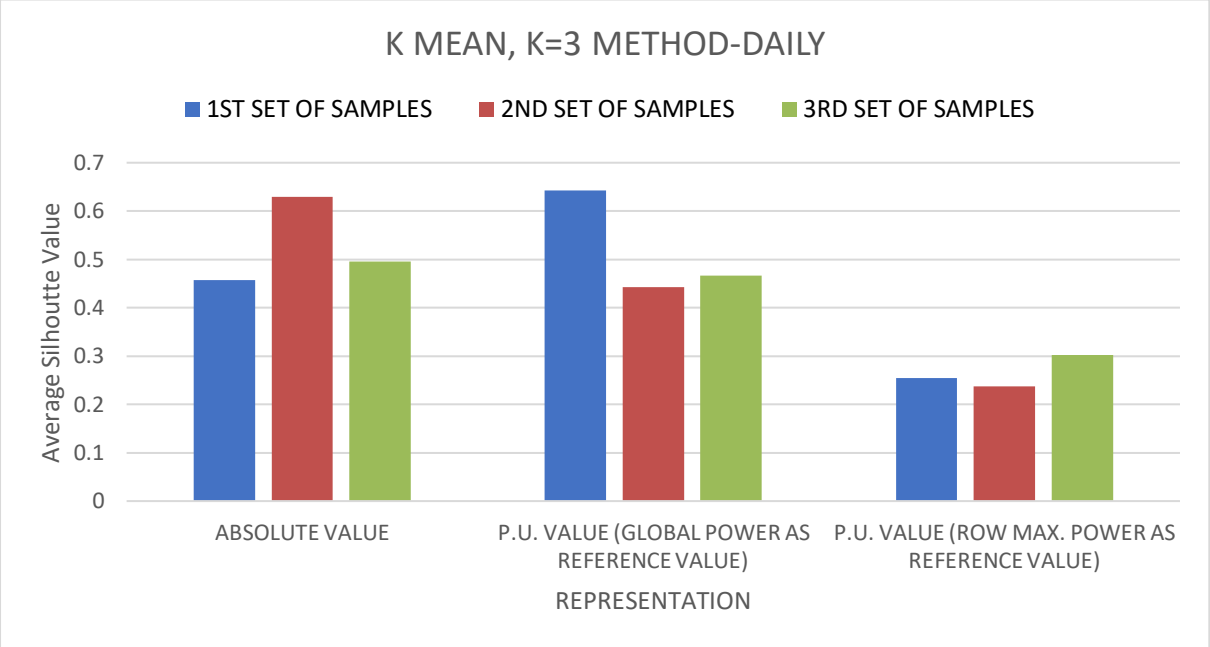


Figure 5.16 Chart of the Average Silhouette value vs the data representation. K-Mean (K=3) method and Daily time step for the three different set of samples are considered.

K-MEAN (K=4 CLUSTERS) METHOD CHARTS

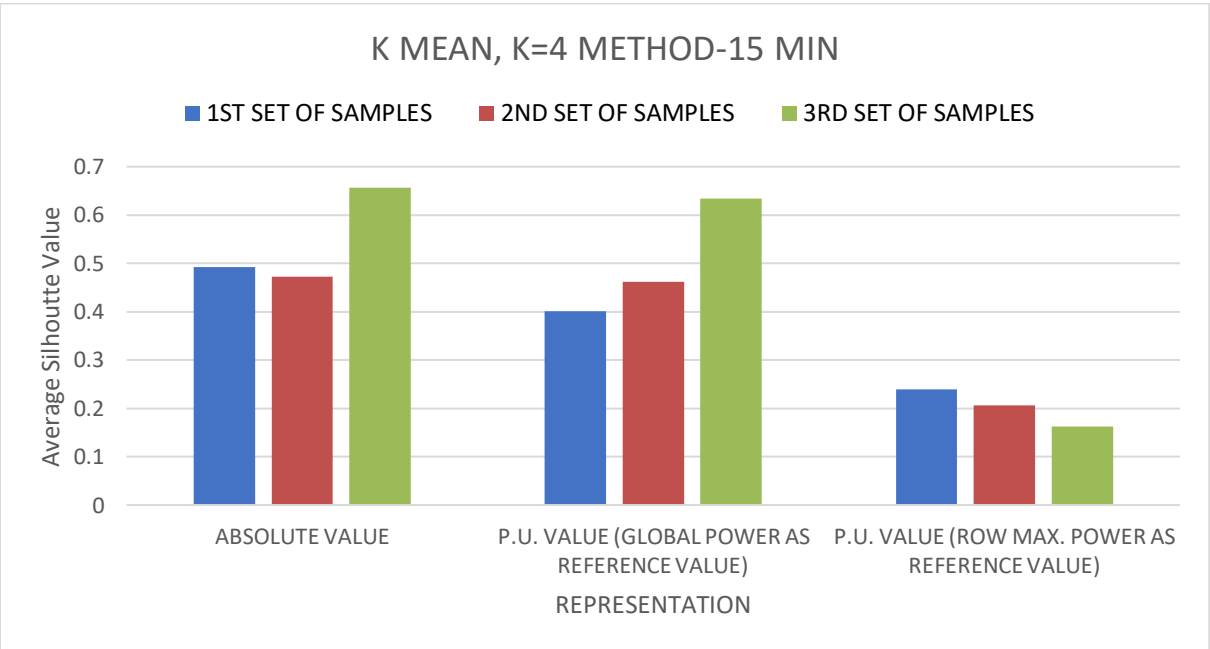


Figure 5.17 Chart of the Average Silhouette value vs the data representation. K-Mean (K=4) method and 15 minutes time step for the three different set of samples are considered.

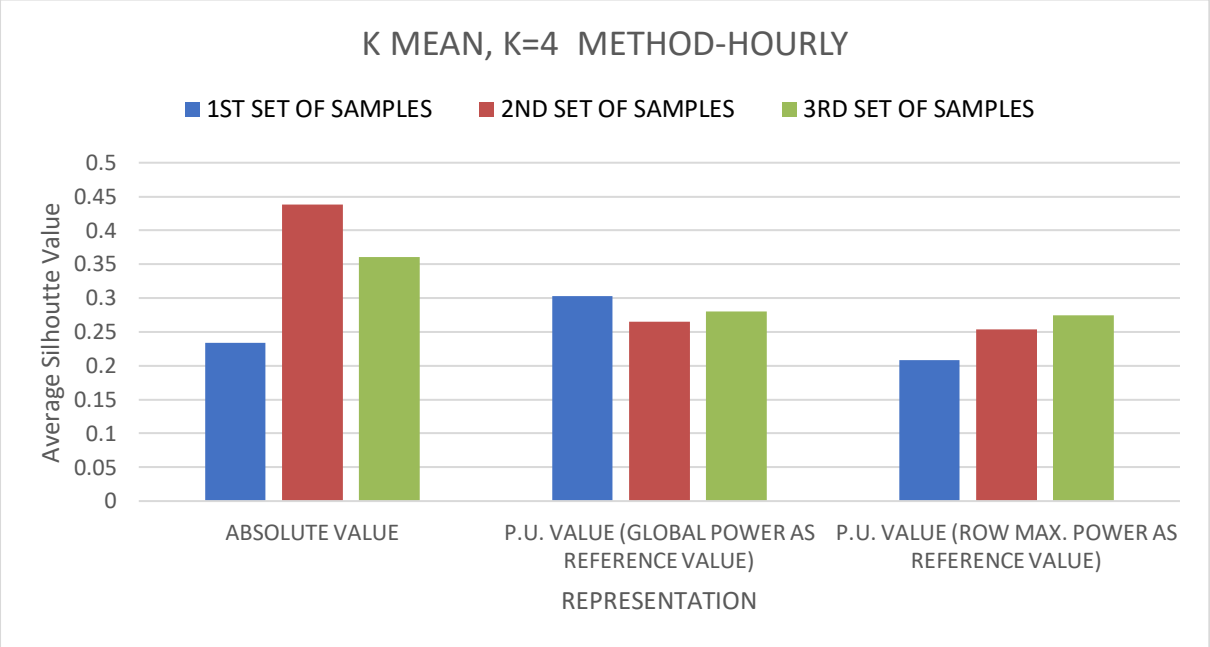


Figure 5.18 Chart of the Average Silhouette value vs the data representation. K-Mean (K=4) method and Hourly time step for the three different set of samples are considered.

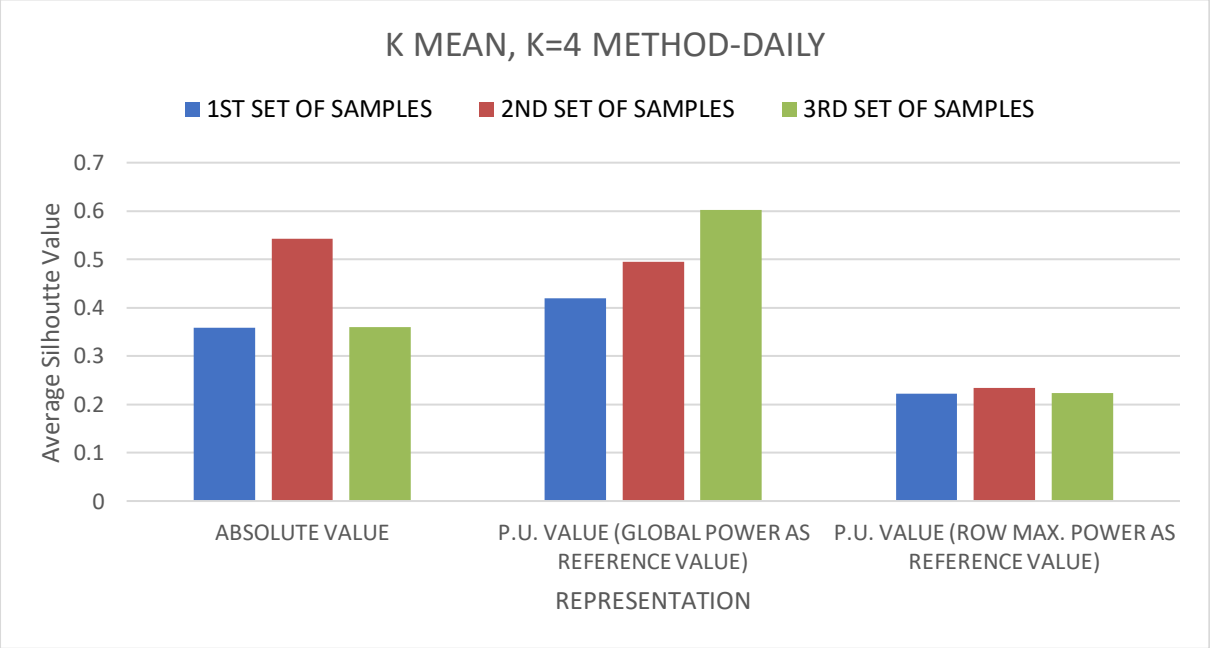


Figure 5.19 Chart of the Average Silhouette value vs the data representation. K-Mean (K=4) method and Daily time step for the three different set of samples are considered.

K-MEAN (K=5 CLUSTERS) METHOD CHARTS

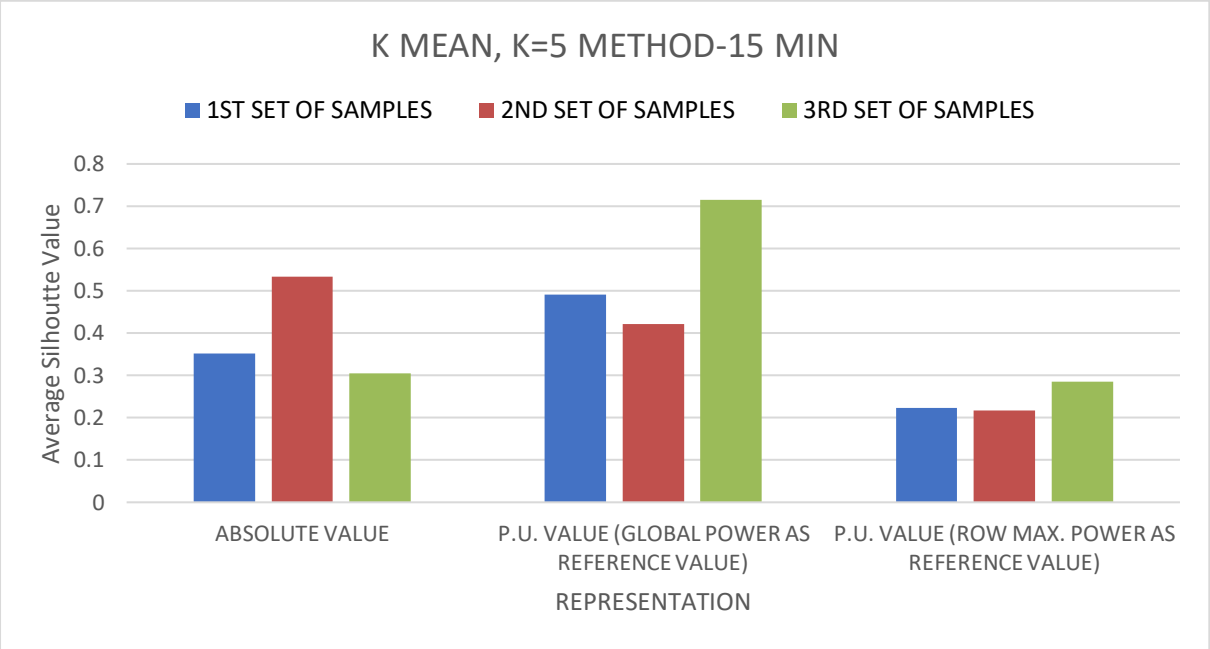


Figure 5.20 Chart of the Average Silhouette value vs the data representation. K-Mean (K=5) method and 15 minutes time step for the three different set of samples are considered.

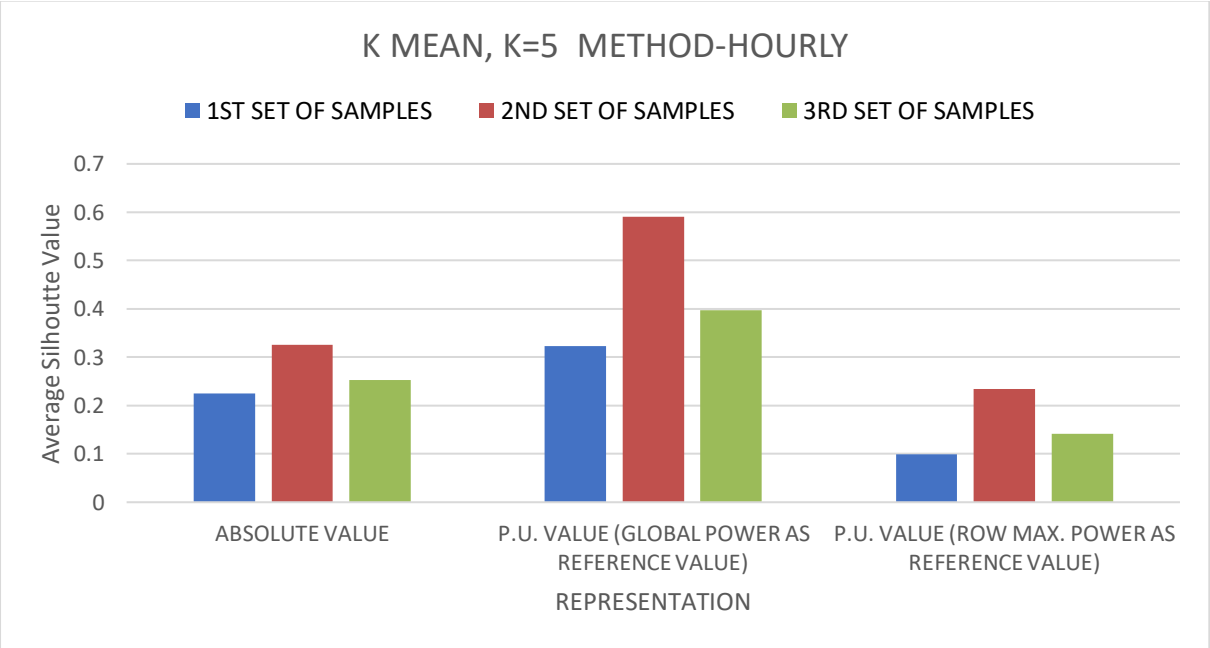


Figure 5.21 Chart of the Average Silhouette value vs the data representation. K-Mean (K=5) method and Hourly time step for the three different set of samples are considered.

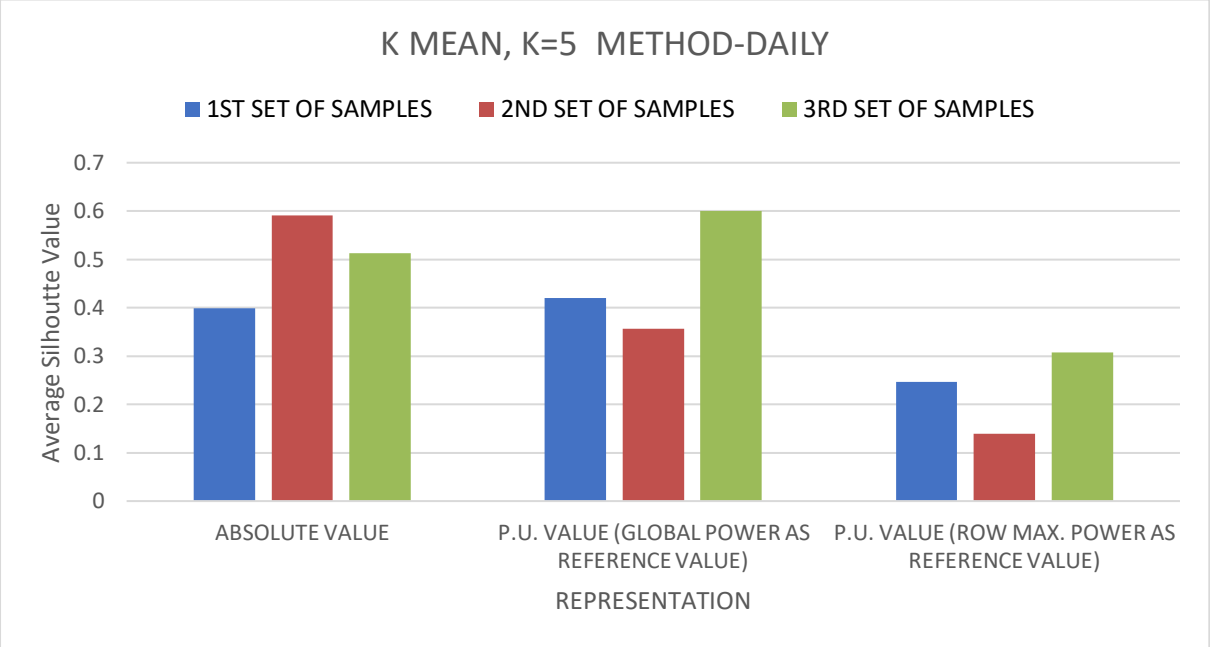


Figure 5.22 Chart of the Average Silhouette value vs the data representation. K-Mean (K=5) method and Daily time step for the three different set of samples are considered.

ANALYSIS ABOUT SAMPLES

About the 1st set of 20 samples, in 16 of 21 charts (more than 75% of the cases) the absolute value representation got the highest average silhouette value in comparison with the other 2 per unit representations. The following representation with highest average silhouette value is the P.U. value based on global maximum power as reference, in 5 of 21 representation. There is not chart for 3rd representation, the P.U. values based on row maximum power as reference with highest average silhouette value.

About the 2nd set of 20 samples, in 17 of 21 charts (almost 80% of the cases) the absolute value representation got the highest average silhouette value in comparison with the other 2 per unit representations. The following representation with highest average silhouette value is the P.U. value based on global maximum power as reference, in 4 of 21 representation. There is not chart using P.U. values based on row maximum power as reference representation with highest average silhouette value.

About the 3rd set of 20 samples, in 17 of 21 (almost 80% of the cases) charts the absolute values representation got the highest average silhouette value in comparison with the other

2 representations. The following representation with highest average silhouette value is the P.U. value based on global maximum power as reference, in 8 of 21 representation. In only one chart, the P.U. values based on row maximum power as reference representation got the highest average silhouette value.

The results show that the absolute representation has better performance compared to the other two, this is because the absolute value representation is presented in its pure state, while the per-unit representations in one case lose very low power data, while the other has a changing reference value for each measuring device.

5.4 Best time step representation

Once the best pattern of input parameters had been selected, a similar process as selecting the best data representation was taken, it will be an analysis to choose the best time-step. Based on the results of three different sets of 20 samples considering the three-time step, it can be said that the best results gotten is when the 15 minutes time step are used. The process of getting the best representation between the three time-step (15 min, hourly and daily time step) was to get the average silhouette value without considering those single member groups with silhouette value equal to 1. Then, for different combination of each method (DBSCAN, HIERARCHICAL, K NEAREST NEIGHBOR, K MEAN-K=2, , K MEAN -K=3, K MEAN -K= 4, K MEAN -K= 5) and each representation (absolute values, and the two P.U. values) some bar charts were made. It means 21 charts were evaluated. In a single chart, the behavior of the three set of samples and the three-time step representation were located (9 bars in one chart). An additional consideration, the analysis for k bigger than 5 were not judged because its average Silhouette value were few lower that the other cases.

The analysis about the results of the three set of samples is shown as following:

The charts and analysis about the results of the three set of samples is shown as following:

DBSCAN METHOD CHARTS

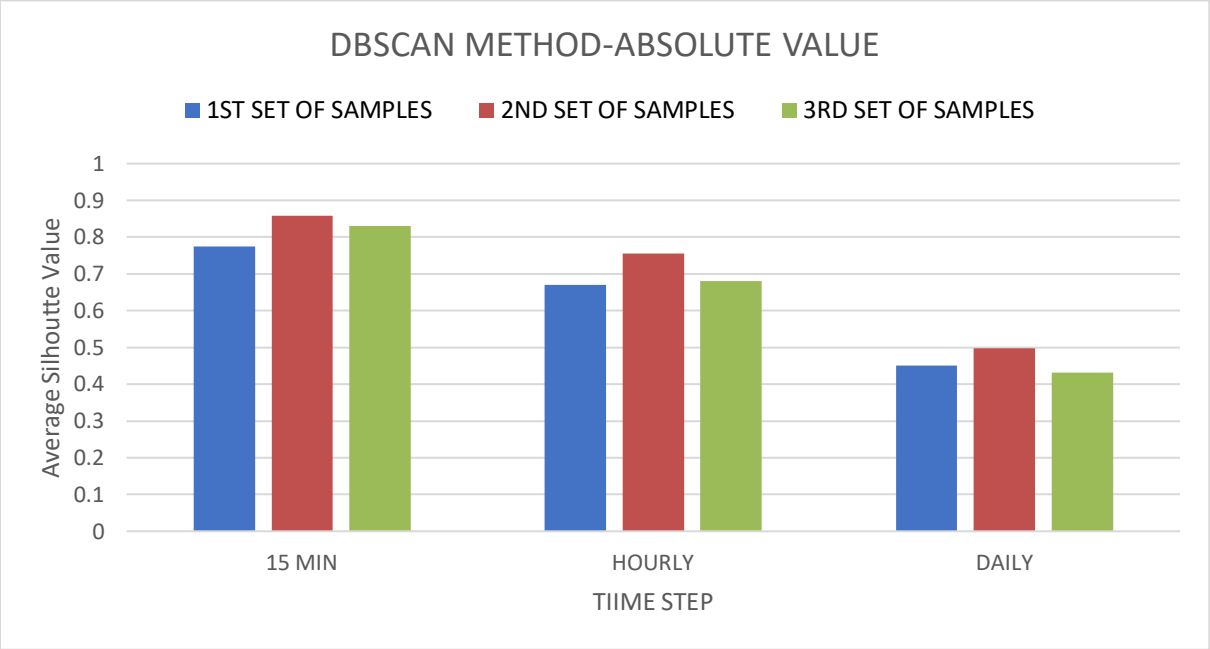


Figure 5.23 Chart of the Average Silhouette value vs the time-step representation. DBSCAN method and Absolute Value representation for the three different set of samples are considered.

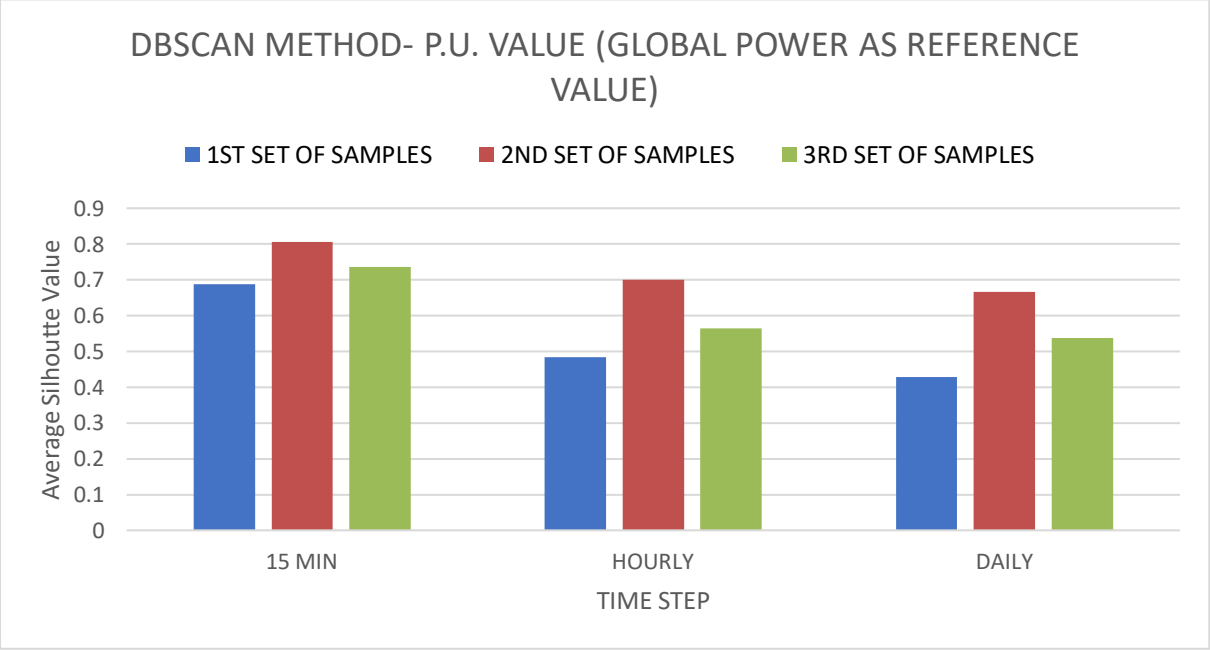


Figure 5.24 Chart of the Average Silhouette value vs the time-step representation. DBSCAN method and p.u. value (Global Power as reference value) for the three different set of samples are considered.

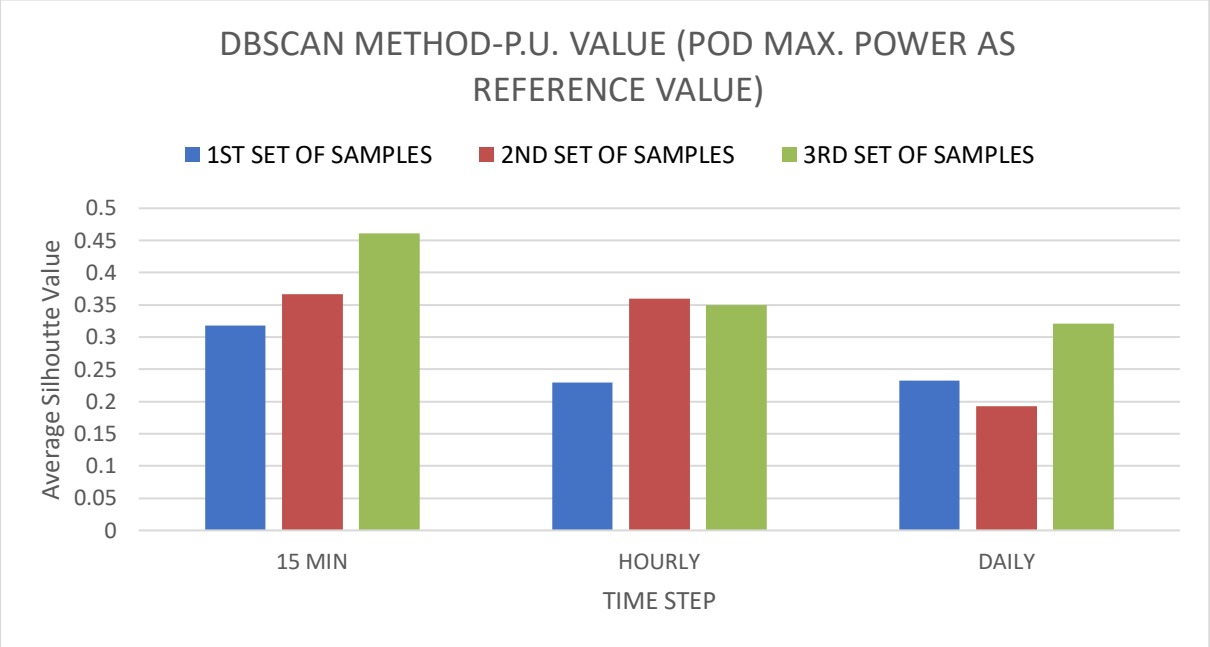


Figure 5.25 Chart of the Average Silhouette value vs the time-step representation. DBSCAN method and p.u. value (POD Max. Power as reference value) for the three different set of samples are considered.

HIERARCHICAL MEHOD CHARTS

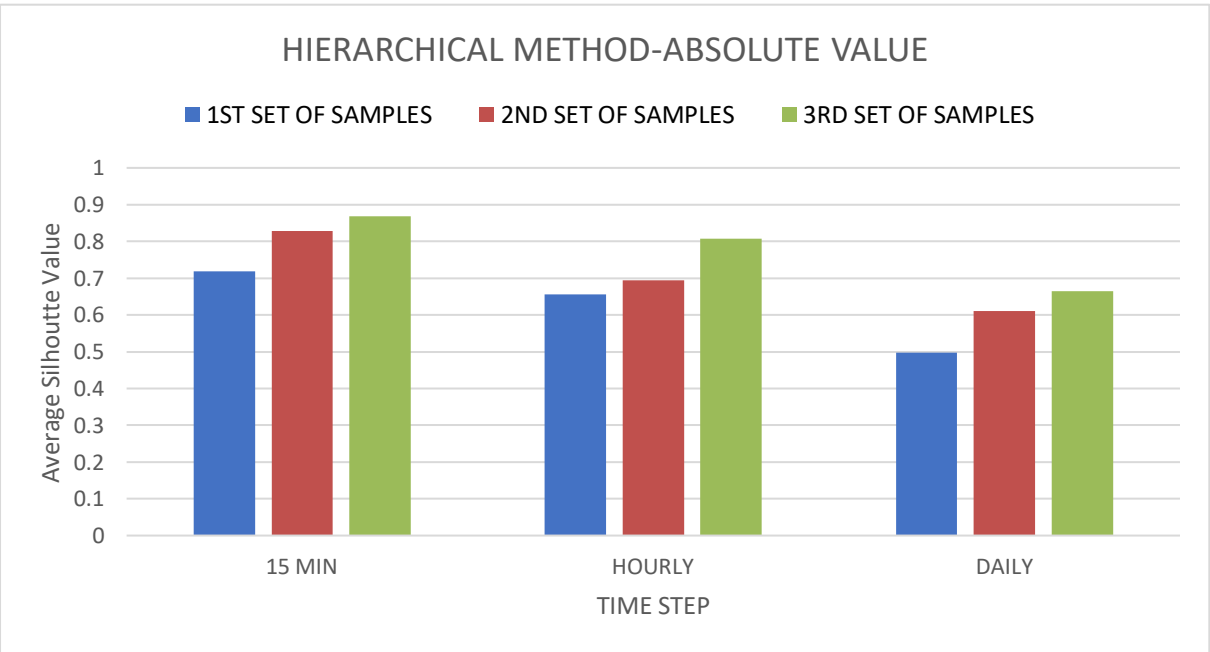


Figure 5.26 Chart of the Average Silhouette value vs the time-step representation. Hierarchical method and Absolute representation for the three different set of samples are considered.

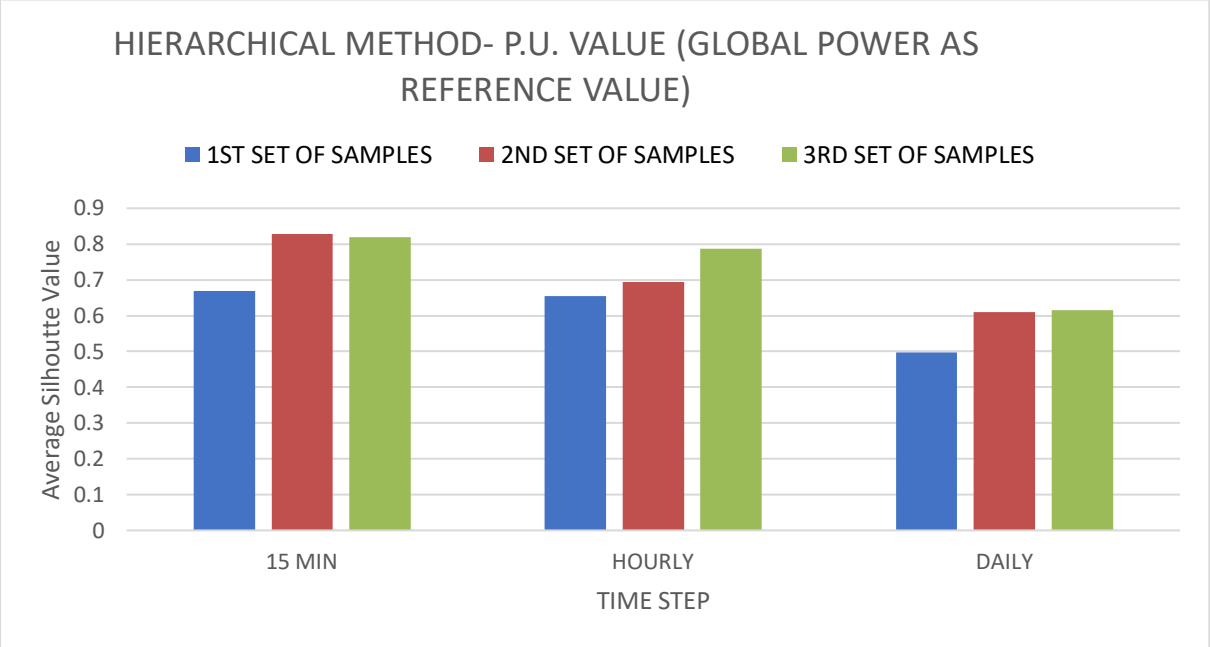


Figure 5.27 Chart of the Average Silhouette value vs the time-step representation. Hierarchical method and p.u. value (Global Power as reference value) for the three different set of samples are considered.

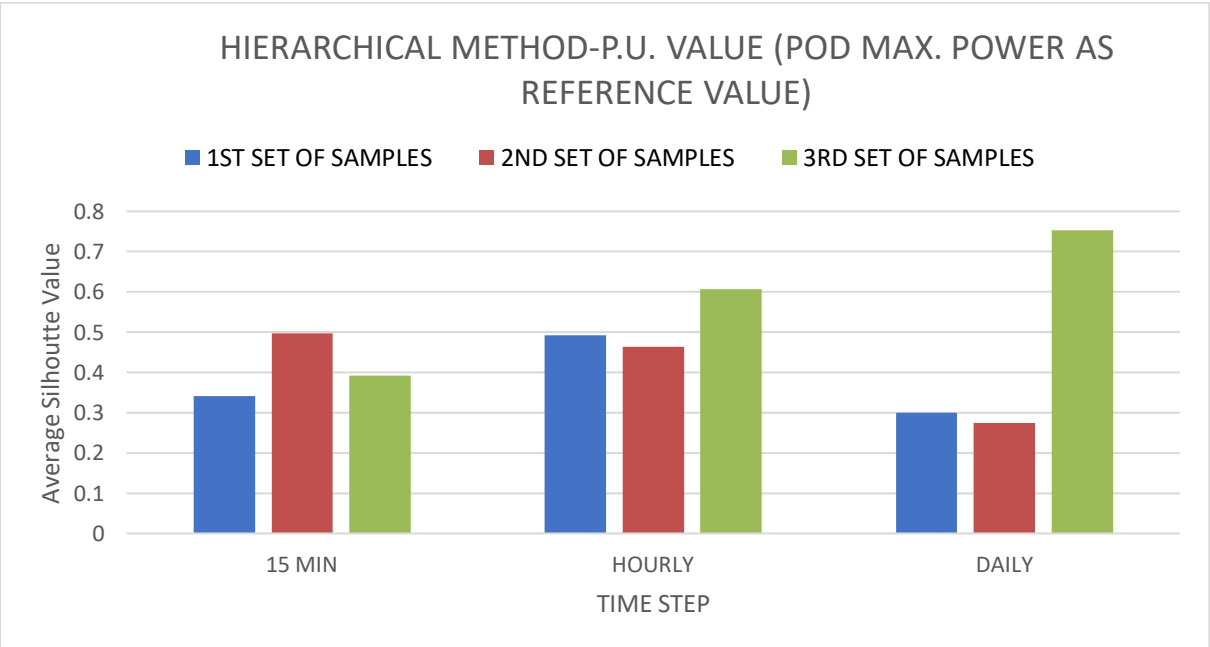


Figure 5.28 Chart of the Average Silhouette value vs the time-step representation. DBSCAN method and p.u. value (POD Max. Power as reference value) for the three different set of samples are considered.

K-NEAREST NEIGHBOR METHOD CHARTS

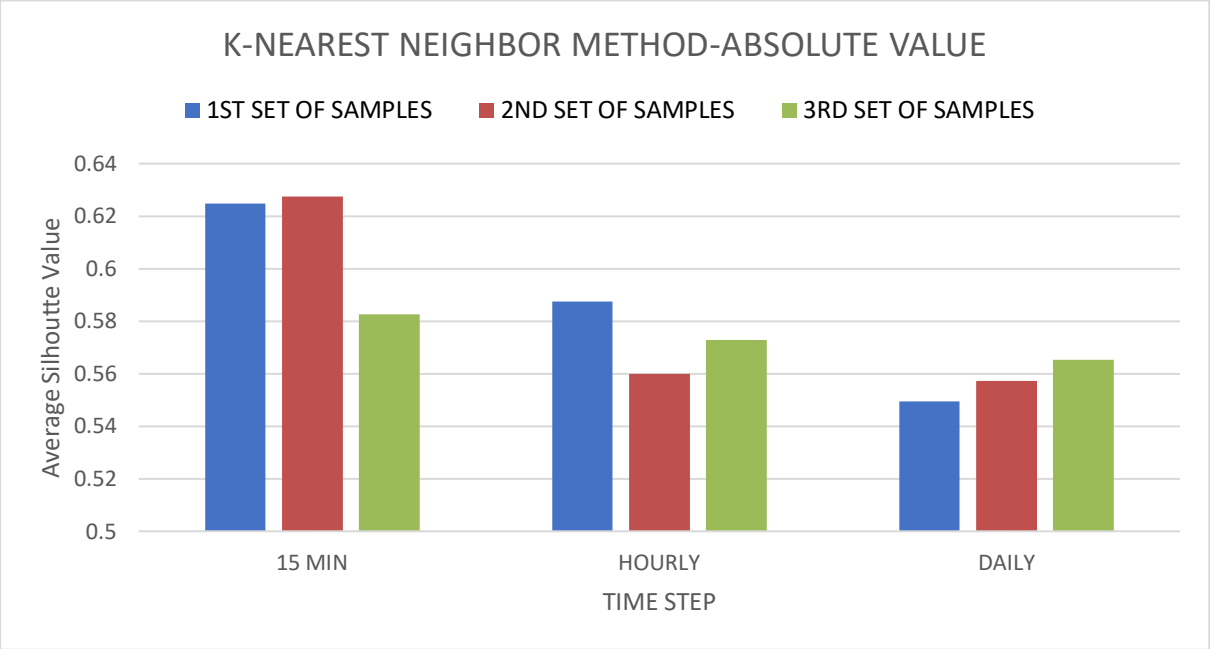


Figure 5.29 Chart of the Average Silhouette value vs the time-step representation. K-Nearest Neighbor method and Absolute representation for the three different set of samples are considered.

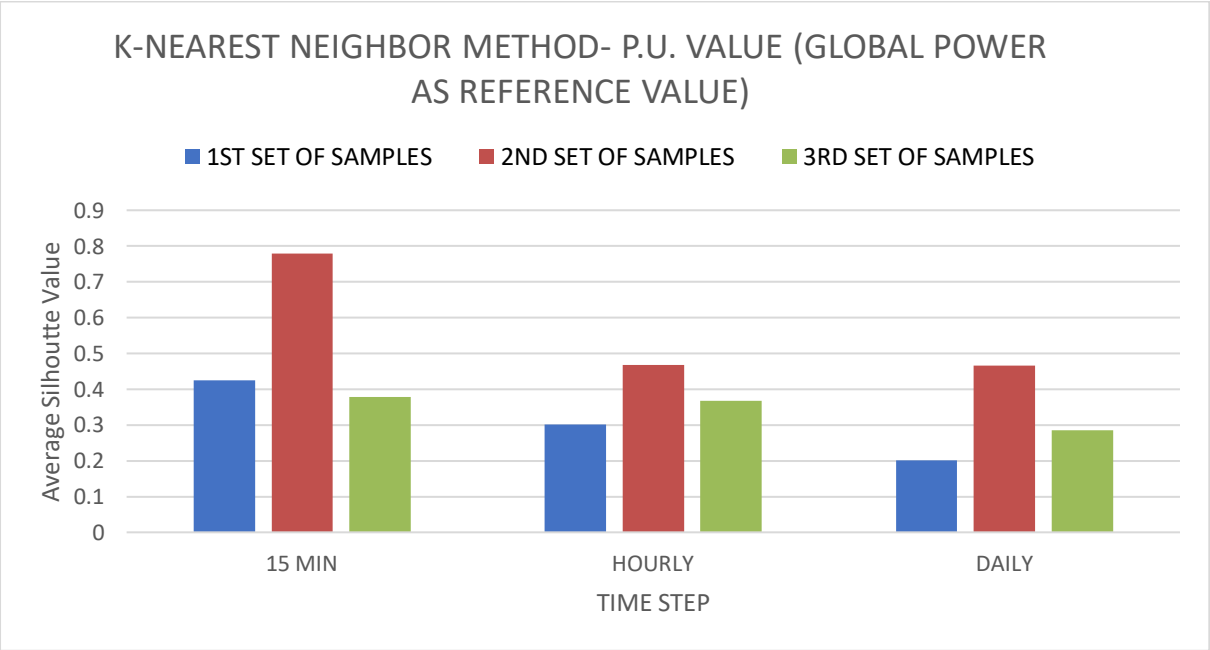


Figure 5.30 Chart of the Average Silhouette value vs the time-step representation. K-Nearest Neighbor method and p.u. value (Global Power as reference value) for the three different set of samples are considered.

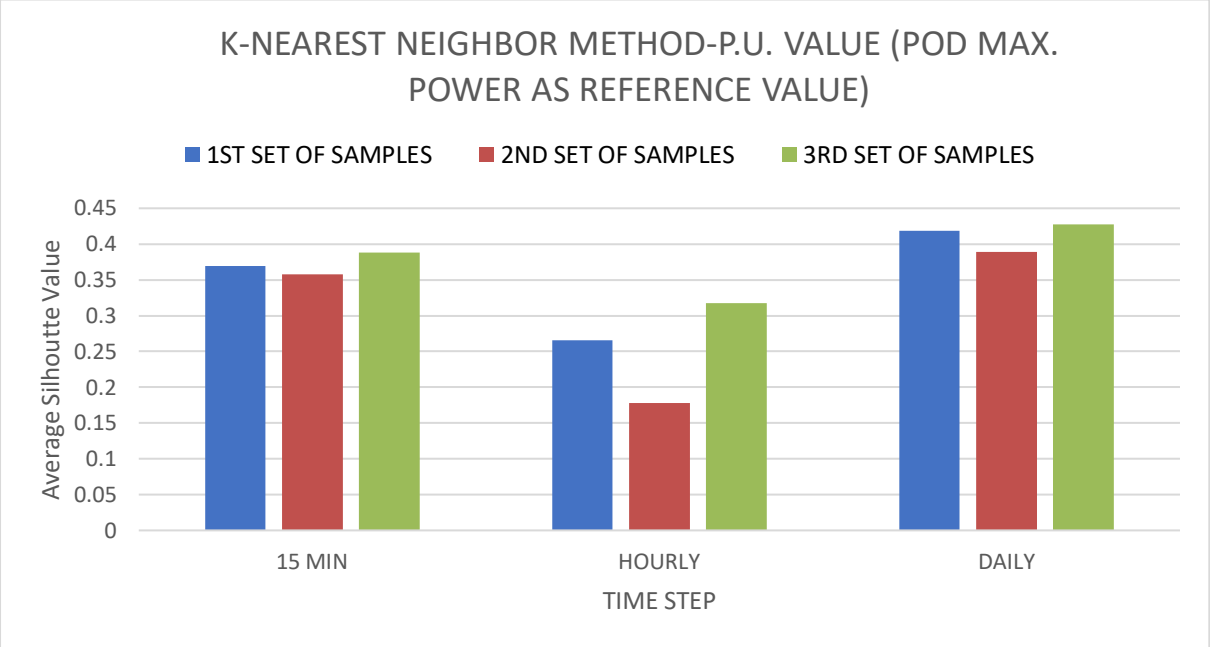


Figure 5.31 Chart of the Average Silhouette value vs the time-step representation. K-Nearest Neighbor method and p.u. value (POD Max. Power as reference value) for the three different set of samples are considered.

K-MEAN (K=2 CLUSTERS) METHOD CHARTS

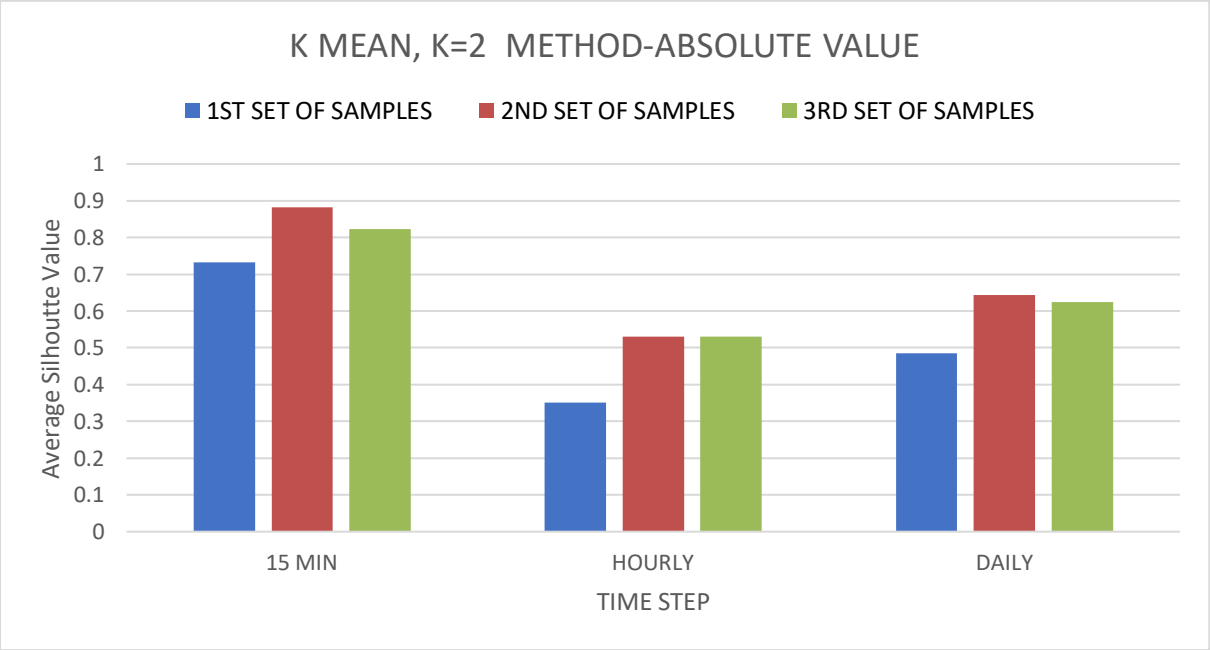


Figure 5.32 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=2 clusters) method and Absolute representation for the three different set of samples are considered.

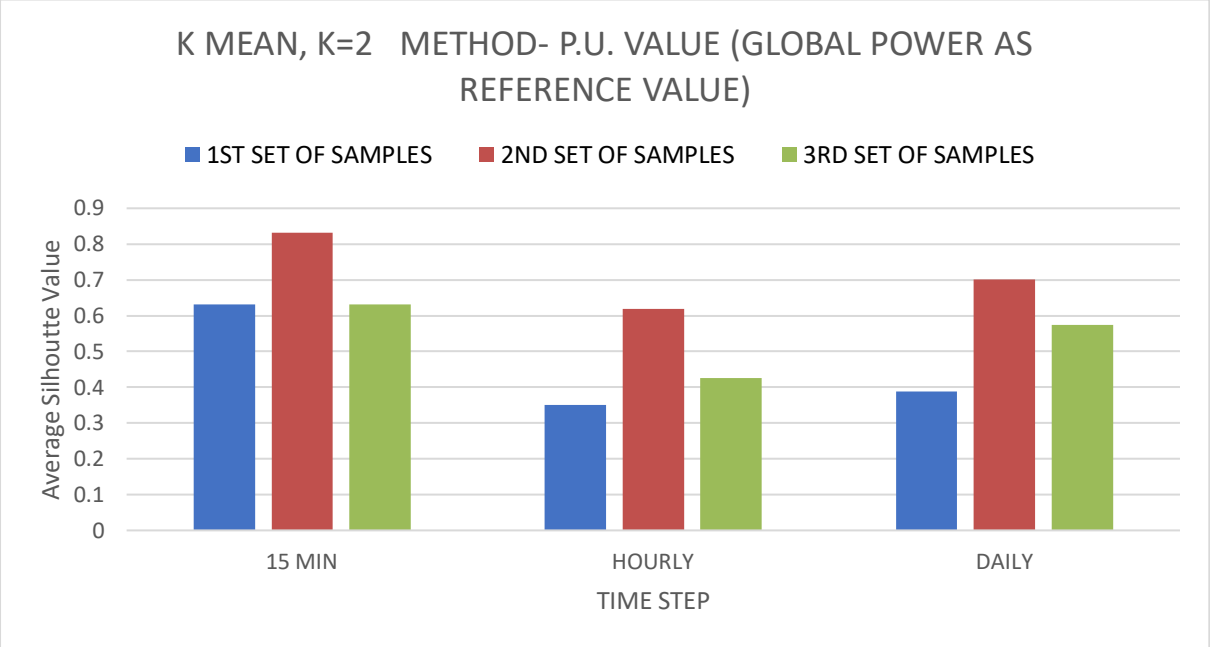


Figure 5.33 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=2 clusters) method and p.u. value (Global Power as reference value) for the three different set of samples are considered.

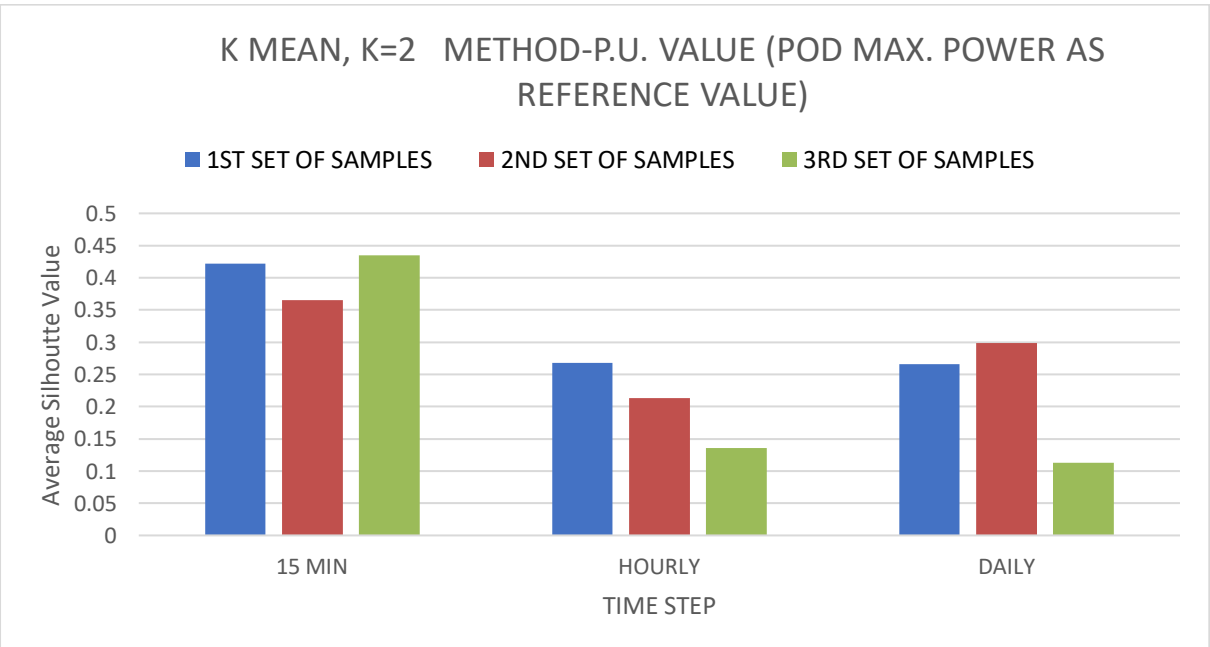


Figure 5.34 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=2 clusters) method and p.u. value (POD Max. Power as reference value) for the three different set of samples are considered.

K-MEAN (K=3 CLUSTERS) METHOD CHARTS

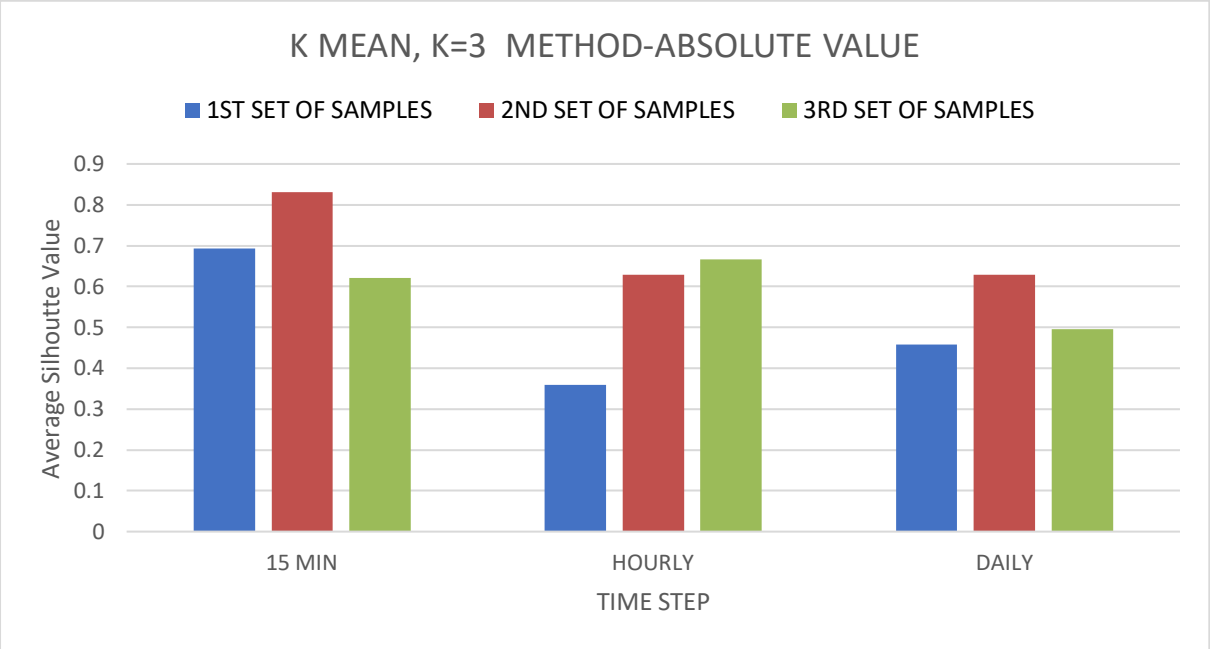


Figure 5.35 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=3 clusters) method and Absolute representation for the three different set of samples are considered.

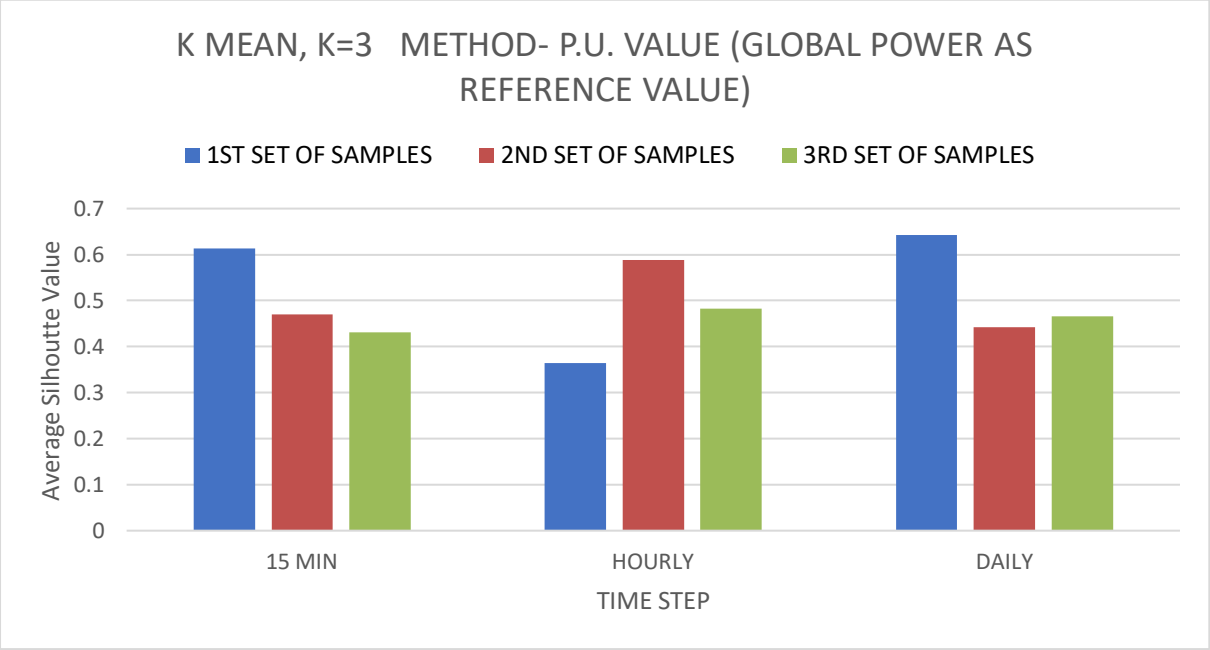


Figure 5.36 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=3 clusters) method and p.u. value (Global Power as reference value) for the three different set of samples are considered.

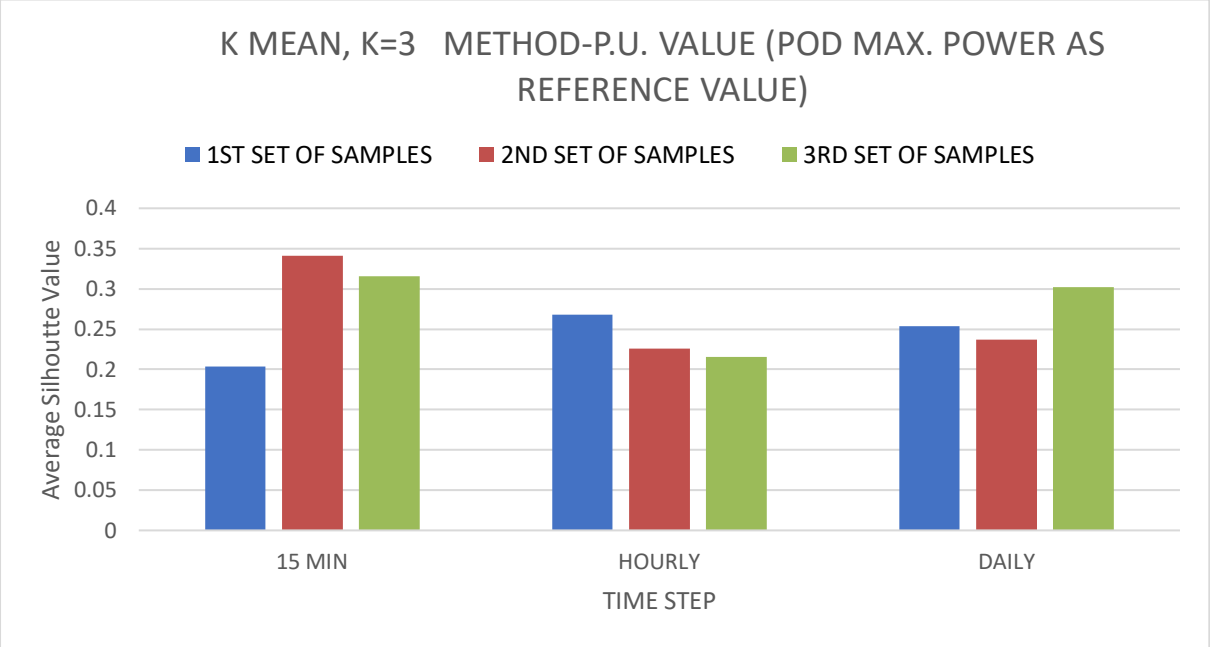


Figure 5.37 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=3 clusters) method and p.u. value (POD Max. Power as reference value) for the three different set of samples are considered.

K-MEAN (K=4 CLUSTERS) METHOD CHARTS

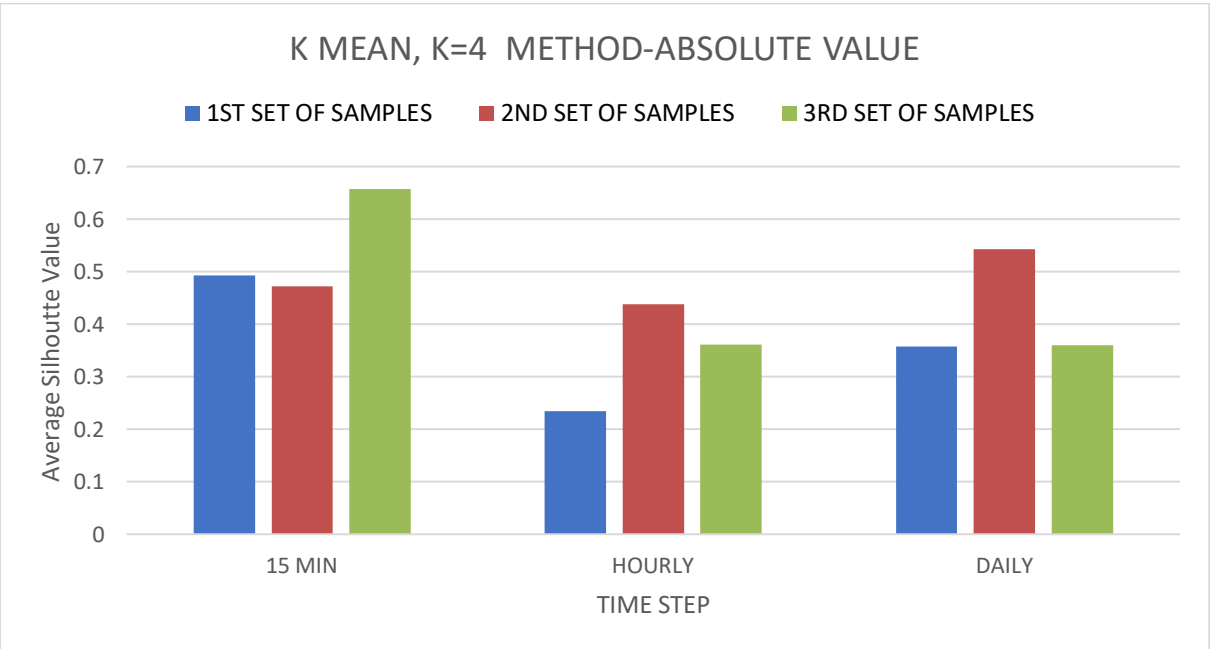


Figure 5.38 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=4 clusters) method and Absolute representation for the three different set of samples are considered.

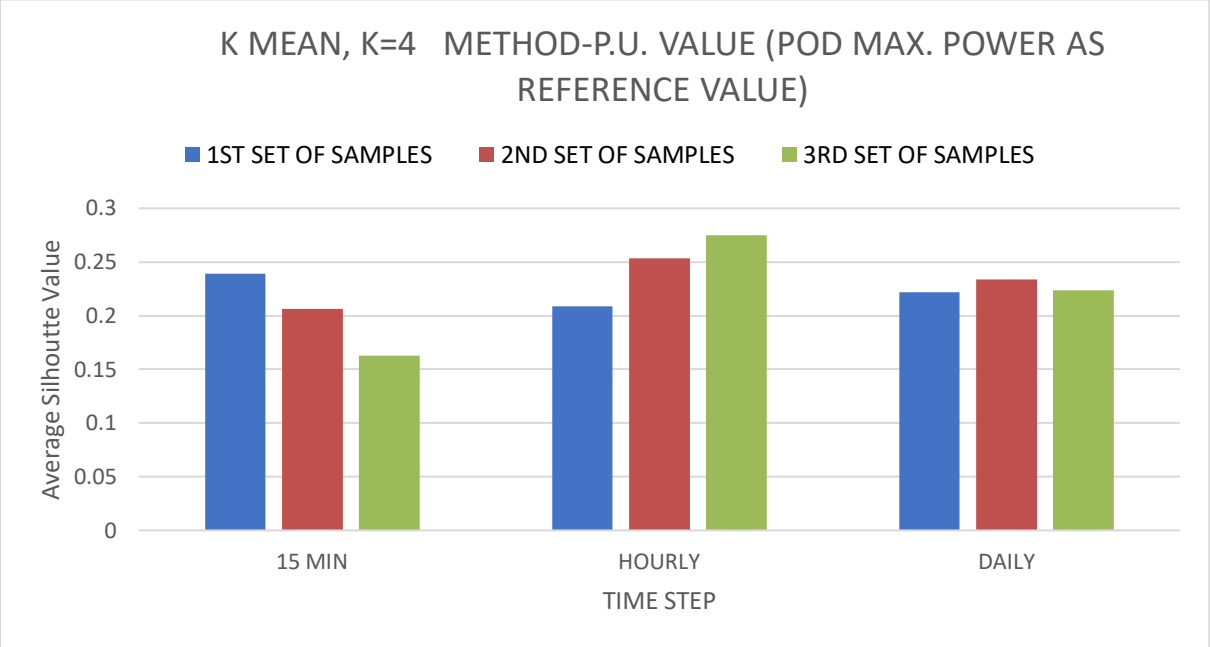


Figure 5.39 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=4 clusters) method and p.u. value (Global Power as reference value) for the three different set of samples are considered.

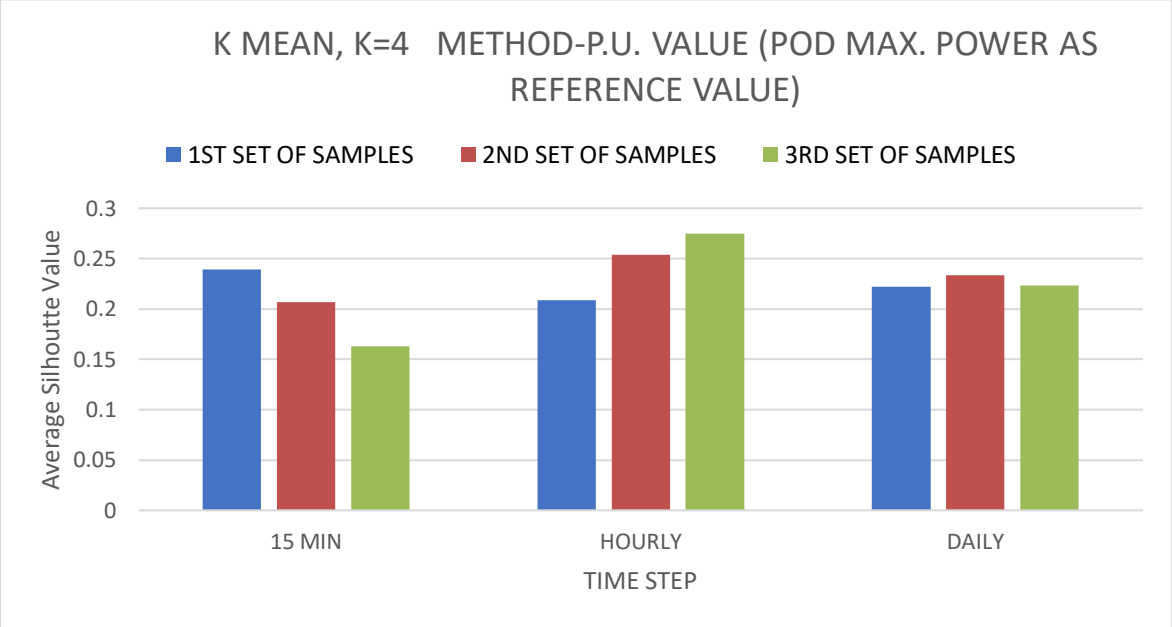


Figure 5.40 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=4 clusters) method and p.u. value (POD Max. Power as reference value) for the three different set of samples are considered.

K-MEAN (K=5 CLUSTERS) METHOD CHARTS

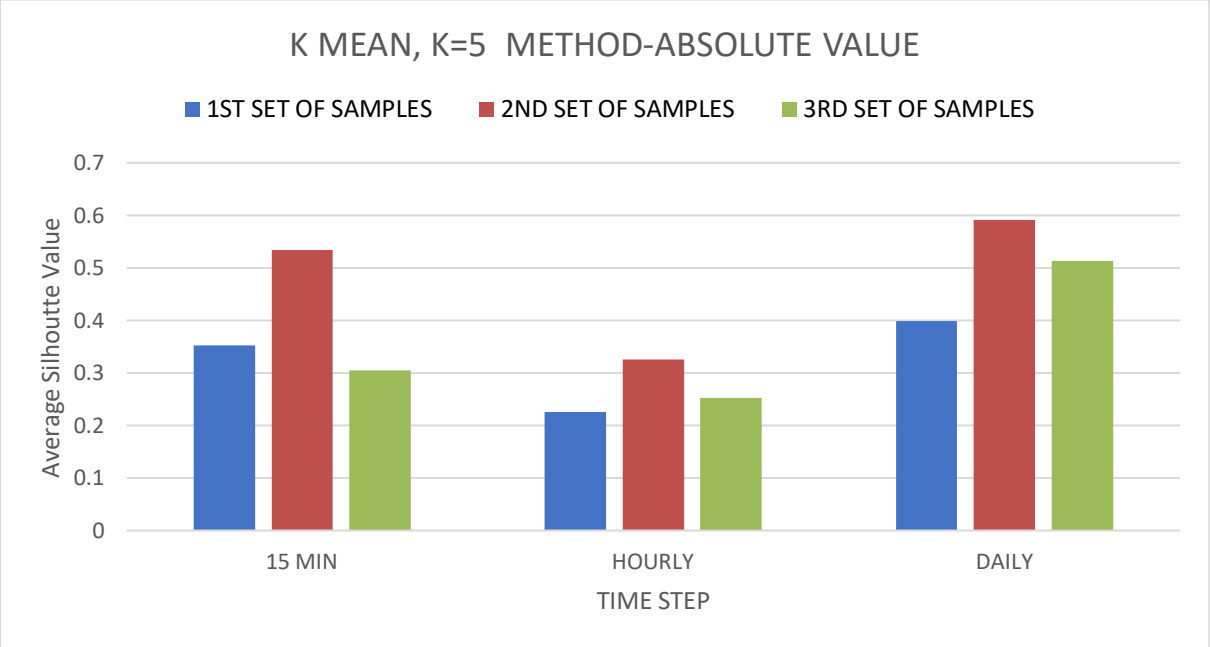


Figure 5.41 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=5 clusters) method and Absolute representation for the three different set of samples are considered.

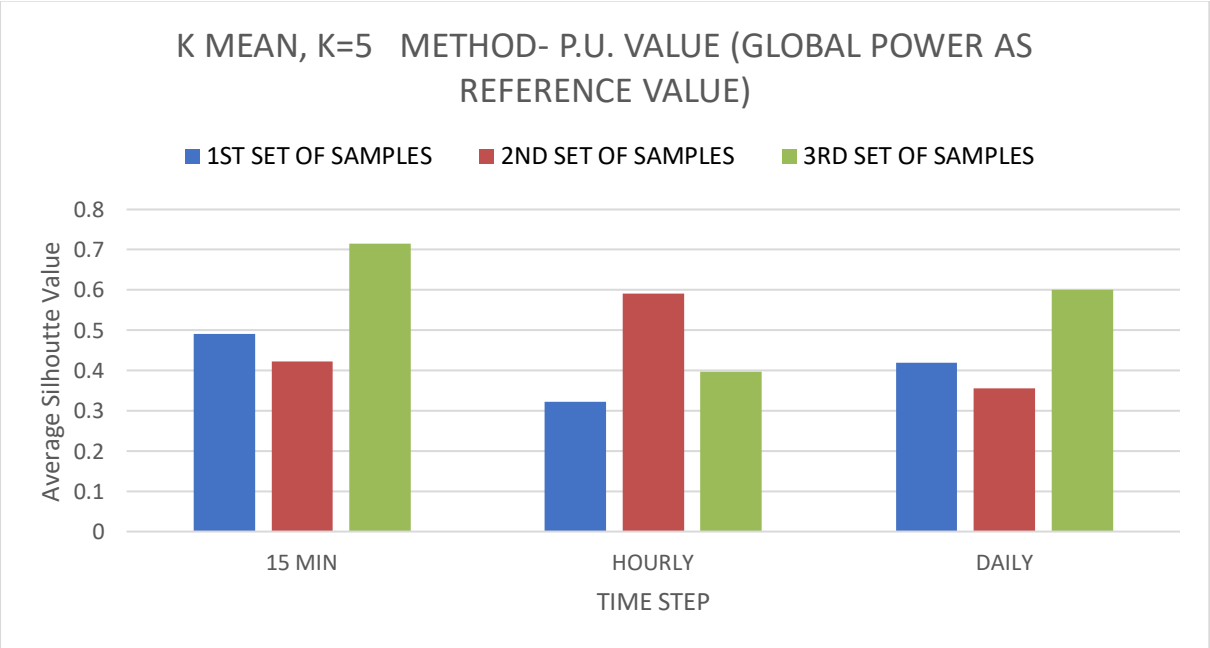


Figure 5.42 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=5 clusters) method and p.u. value (Global Power as reference value) for the three different set of samples are considered.

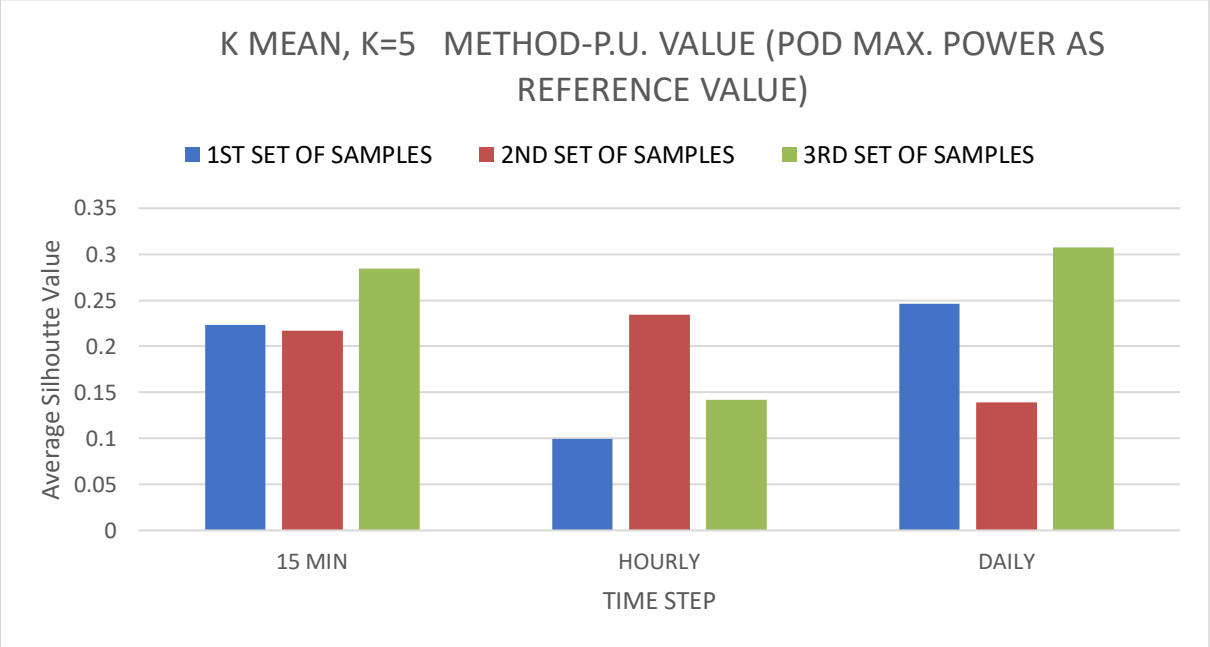


Figure 5.43 Chart of the Average Silhouette value vs the time-step representation. K-Mean (K=5 clusters) method and p.u. value (POD Max. Power as reference value) for the three different set of samples are considered.

ANALYSIS ABOUT SAMPLES

About the 1st set of 20 samples, in 17 of 21 charts (more than 80% of the cases) the 15 min time-step representation got the highest average silhouette value in comparison with the other time step representations. The following representation with highest average silhouette value is hourly time representation, in 2 of 21 representation. There are 2 charts with highest average silhouette value where the daily time representation is considered.

About the 2nd set of 20 samples, in 18 of 21 charts (more than 85% of the cases) the 15 min time-step representation got the highest average silhouette value in comparison with the other time representations. The following representation with highest average silhouette value is hourly time representation, in 3 of 21 representation. There are only charts with highest average silhouette value where the daily time representation is considered.

About the 3rd set of 20 samples, in 17 of 21 charts (more than 80% of the cases) the 15 min time-step representation got the highest average silhouette value in comparison with the other time representations. The following representation with highest average silhouette value is hourly time representation, in 2 of 21 representation. There are 2 charts with highest average silhouette value where the daily time representation is considered

The results show that the 15 minutes representation has better performance compared to the other two, this is because the more spaced we have the measurements (every 15 minutes instead of every hour or every 24 hours), the more accurate our analysis will be.

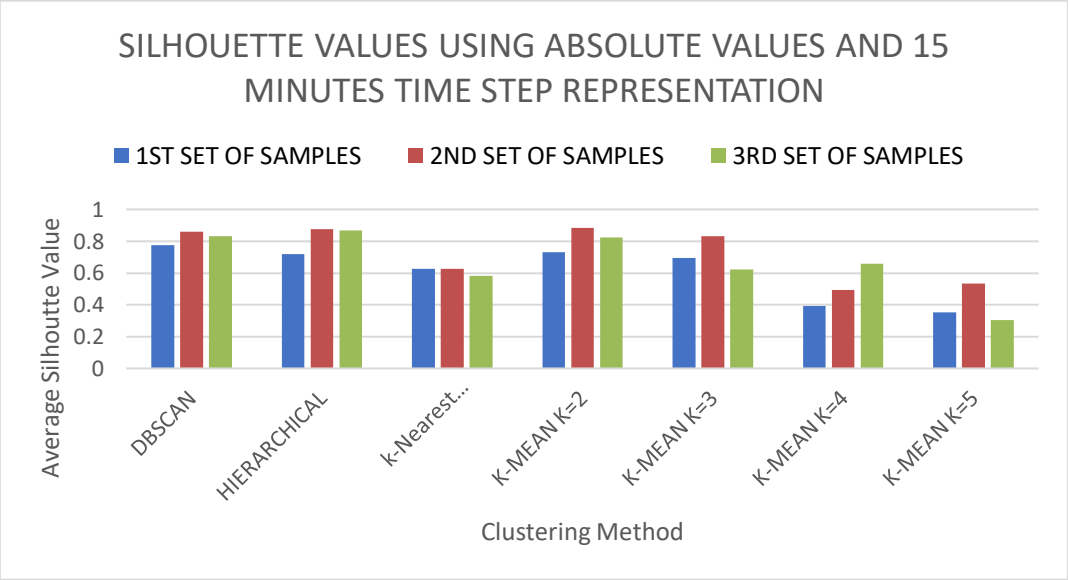


Figure 5.44 Chart of the average Silhouette value vs clustering method, considering absolute values and a time-step of 15 minutes. The three set of samples are being analyzed.

5.5 Performance of the methods considering the Samples

Table 5.1 Table of the average Silhouette value for different clustering method, considering absolute values and a time-step of 15 minutes. The three set of samples are being analyzed.

SILHOUETTE VALUES USING ABSOLUTE VALUES AND 15 MINUTES TIME STEP REPRESENTATION			
METHOD/SET	**1ST**	**2ND**	**3RD**
DBSCAN	0.774924	0.858394	0.829776
HIERARCHICAL	0.718508	0.877712	0.868859
k-Nearest Neighbor	0.624771	0.627584	0.582493
K-MEAN K=2	0.731829	0.88202	0.823388
K-MEAN K=3	0.692888	0.830584	0.621562
K-MEAN K=4	0.392558	0.492558	0.65706
K-MEAN K=5	0.352277	0.533605	0.304814

According to the results which involve the representation and time step selection, it can be said that the arrangement to get the best results is when the absolute values and 15 minutes time step are considered.

Once the best combination is gotten, the next step is select the best approach. The three different set of samples randomly chosen were analyzed considering the absolute values with 15 minutes time-step. The charts results of each set of samples are shown in Fig. 5.43.

For the 1st set of samples, in the unsupervised side the Hierarchical method has the highest average Silhouette value, 0.774924. On the supervised side, the K MEAN method with k=2, has the highest clustering value, 0.731829 closely followed by the K MEAN method with k=3, 0.692888.

For the 2nd set of samples, in the unsupervised side the Hierarchical method has the highest average Silhouette value, 0.877712. On the supervised side, the K MEAN method with

k=2, has the highest clustering value, 0.88202 closely followed by the K MEAN method with k=3, 0.830584

For the 3rd set of samples, in the unsupervised side the Hierarchical method has the highest average Silhouette value, 0.868859. On the supervised side, the K MEAN method with k=2, has the highest clustering value, 0.823388 followed by the K MEAN method with k=3, 0.62156.

In summary, there are two methods with the best performance for the three sets of samples. For the unsupervised side, The Hierarchical method for the three set of samples, has the highest average silhouette value (without counting those single member groups with average silhouette value equal to 1). From the supervised side, the K MEAN method (with k=2 clusters) for the three set of samples, has the highest average silhouette value (without counting those single member groups with average silhouette value equal to 1).

5.6 Sensitivity Analysis of the Whole Data

As it was done for the sample data, a sensitivity analysis was run for the whole data. The input parameters were changed until a maximization of the Silhouette value was gotten.

As it happened in the sample analysis, those one single member groups with a Silhouette value equal to 1 were set aside of the evaluation.

Many graphs of the results are shown so that the reader can get an idea of the behavior of the silhouette values for each method analyzed, depending the input parameter.

HIERARCHICAL METHOD

BUFFER SIZE “d” FROM 1 TO 20

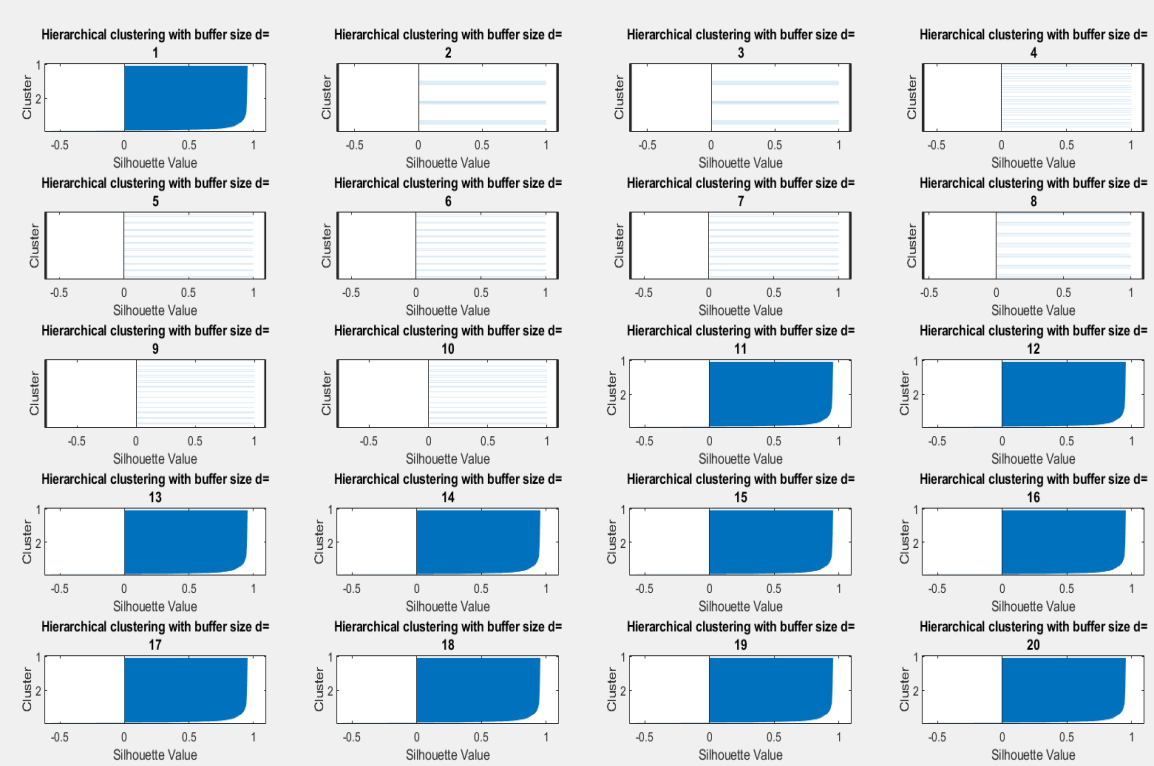


Figure 5.45 Charts of the average Silhouette value vs different buffer size “d”. Hierarchical method is considered.

As it can be seen, the best results on Silhouette values are given when the variable “d”, which represents the size of the buffer is more than 11. It can be viewed that those charts have the same Silhouette value equal to 0.92. But it can be seen in these cases the Hierarchical method, gives the best result with two groups where one of them is a single element.

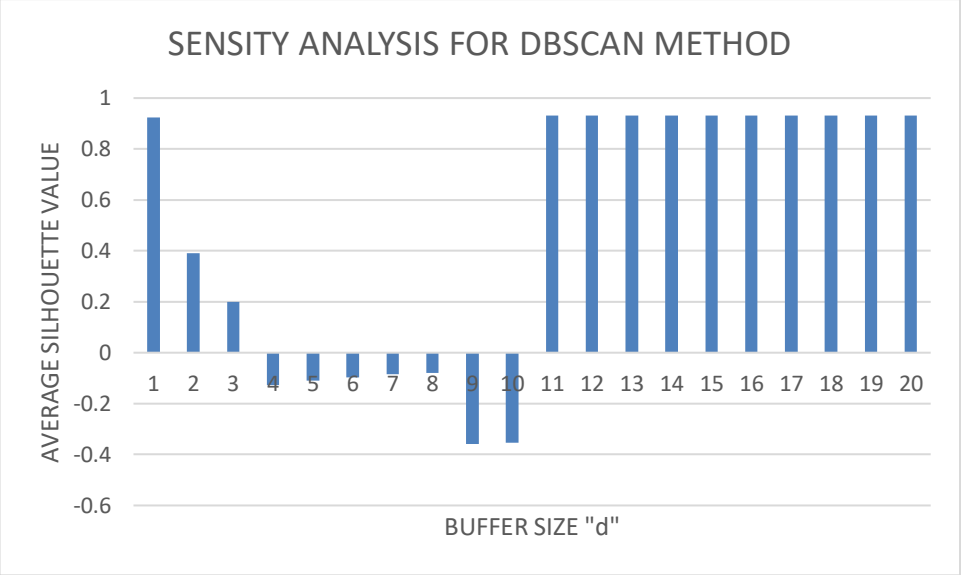


Figure 5.46 Chart of the average Silhouette value vs different buffer size d. Hierarchical method is considered.

DBSCAN METHOD

MIN POINTS=10 AND EPSILON FROM 5000 TO 24000 (STEPS OF 1000)

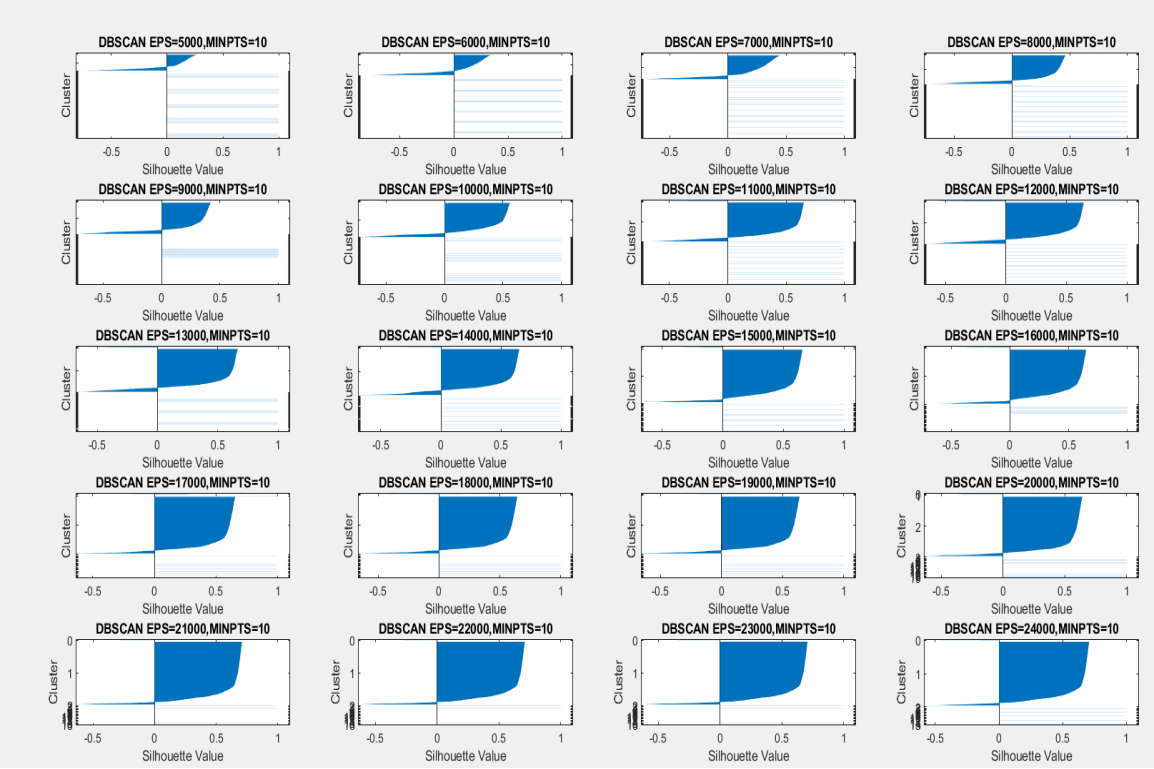


Figure 5.47 Charts of the average Silhouette value vs different values of the input parameter (epsilon=10 and different values of Minimum Points). DBSCAN method is considered.

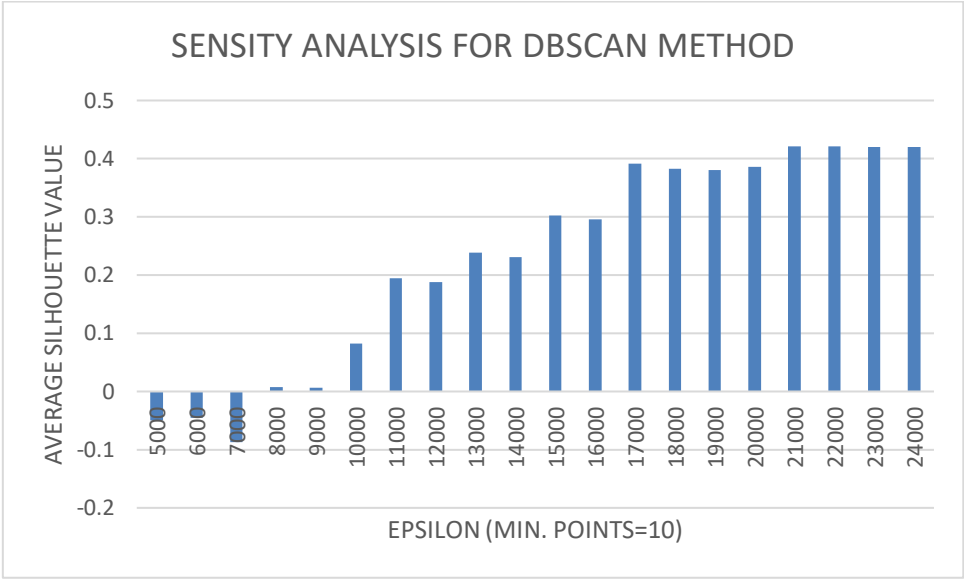


Figure 5.48 Chart of the average Silhouette values different values of the input parameter (epsilon=10 and different values of Minimum Points). DBSCAN method is considered.

MIN POINTS=8 AND EPSILON FROM 5000 TO 24000 (STEPS OF 1000)

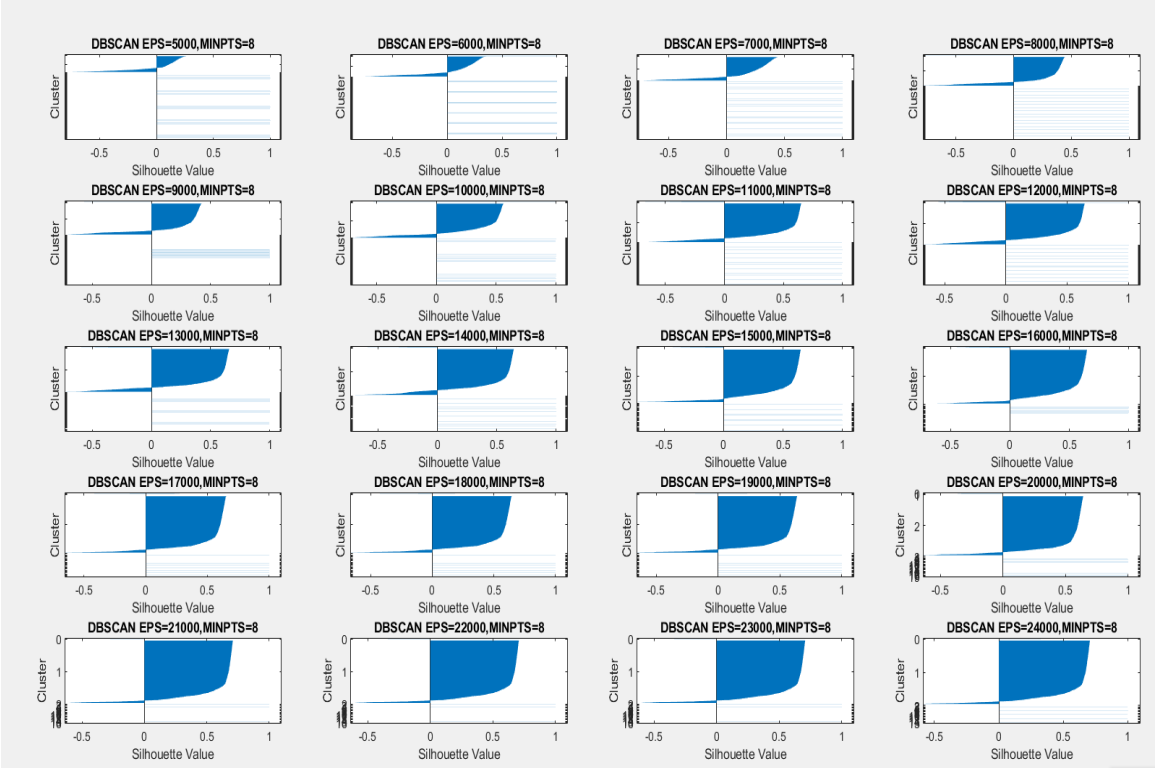


Figure 5.49 Charts of the average Silhouette value vs different values of the input parameter (epsilon=8 and different values of Minimum Points). DBSCAN method is considered.

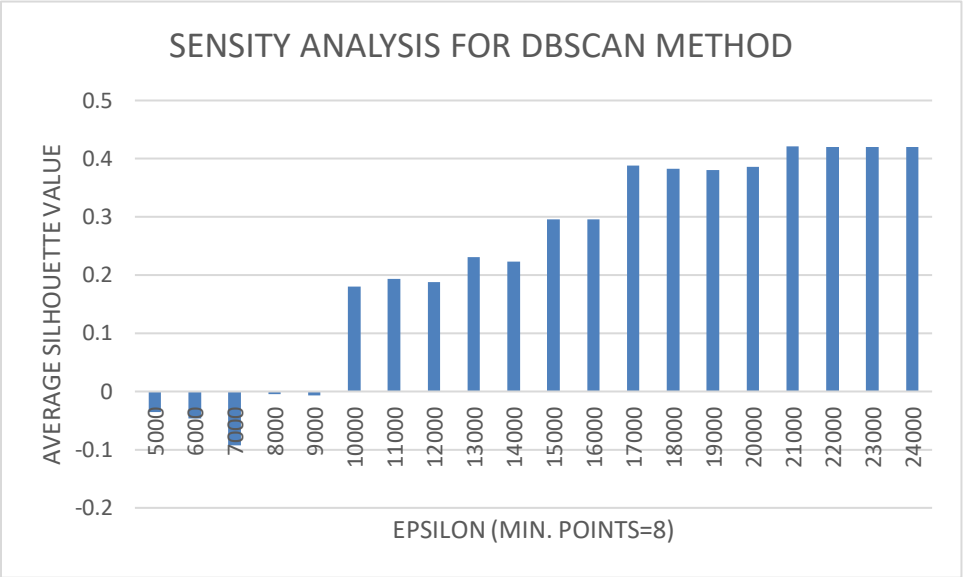


Figure 5.50 Chart of the average Silhouette values different values of the input parameter (epsilon=8 and different values of Minimum Points). DBSCAN method is considered.

MIN POINTS=6 AND EPSILON FROM 5000 TO 24000 (STEPS OF 1000)

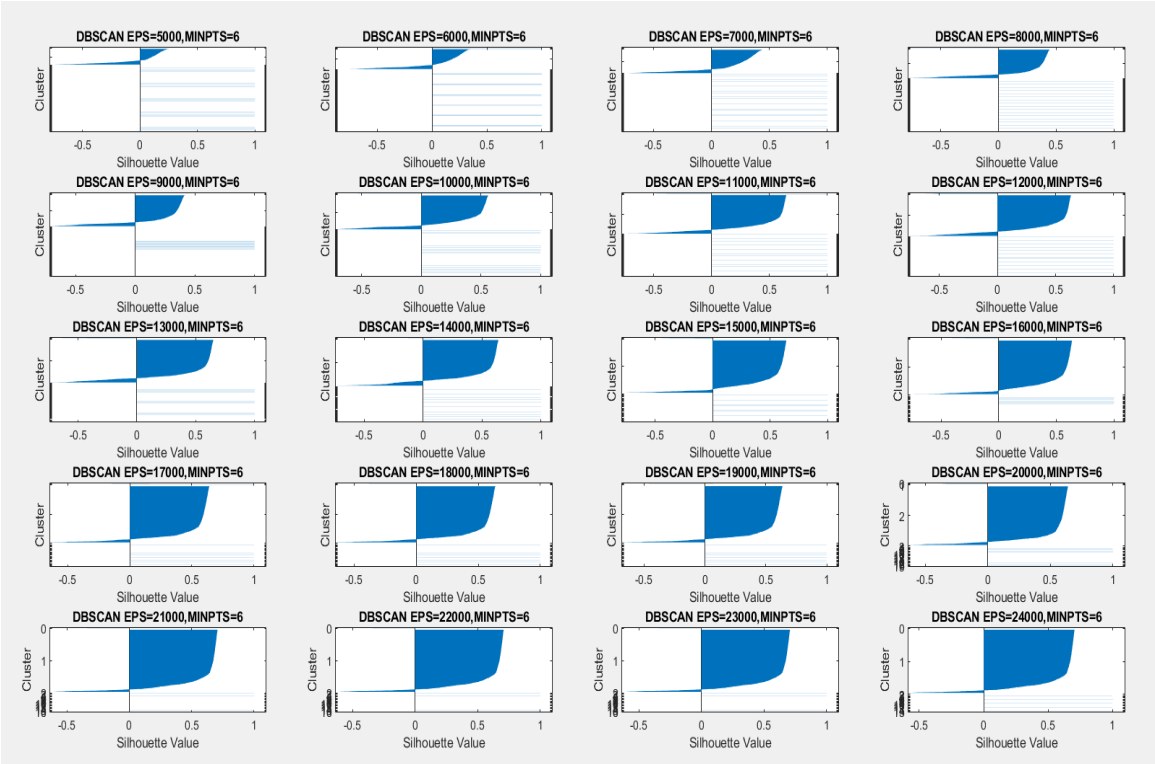


Figure 5.51 Charts of the average Silhouette value vs different values of the input parameter (epsilon=6 and different values of Minimum Points). DBSCAN method is considered.

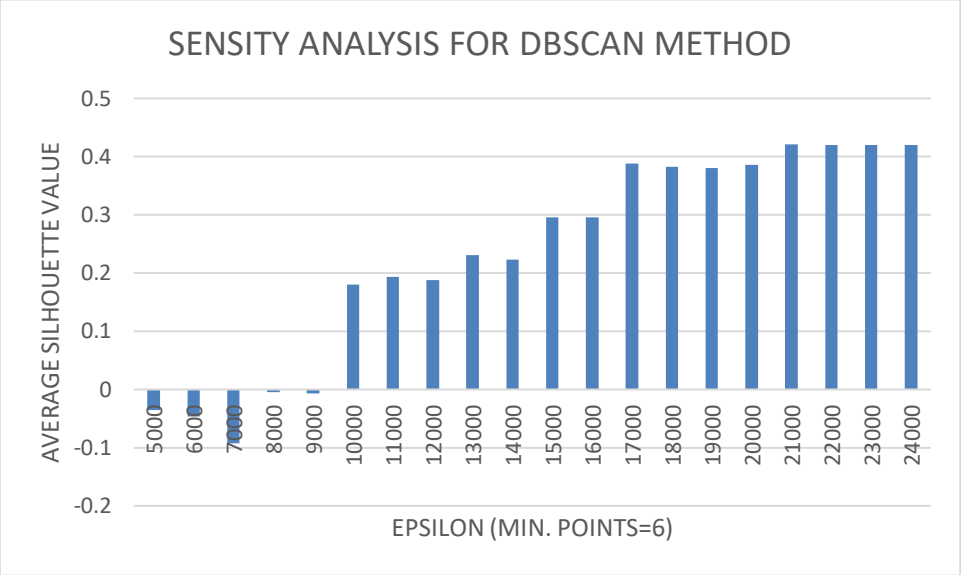


Figure 5.52 Chart of the average Silhouette values different values of the input parameter (epsilon=6 and different values of Minimum Points). DBSCAN method is considered.

It was observed that the silhouette values in all the cases analyzed do not exceed 0.45. And even in many cases, there are some negative Silhouette values. This can give us an idea about the low performance of DBSCAN method.

K NEAREST NEIGHBOR FROM K= 1 TO K= 20

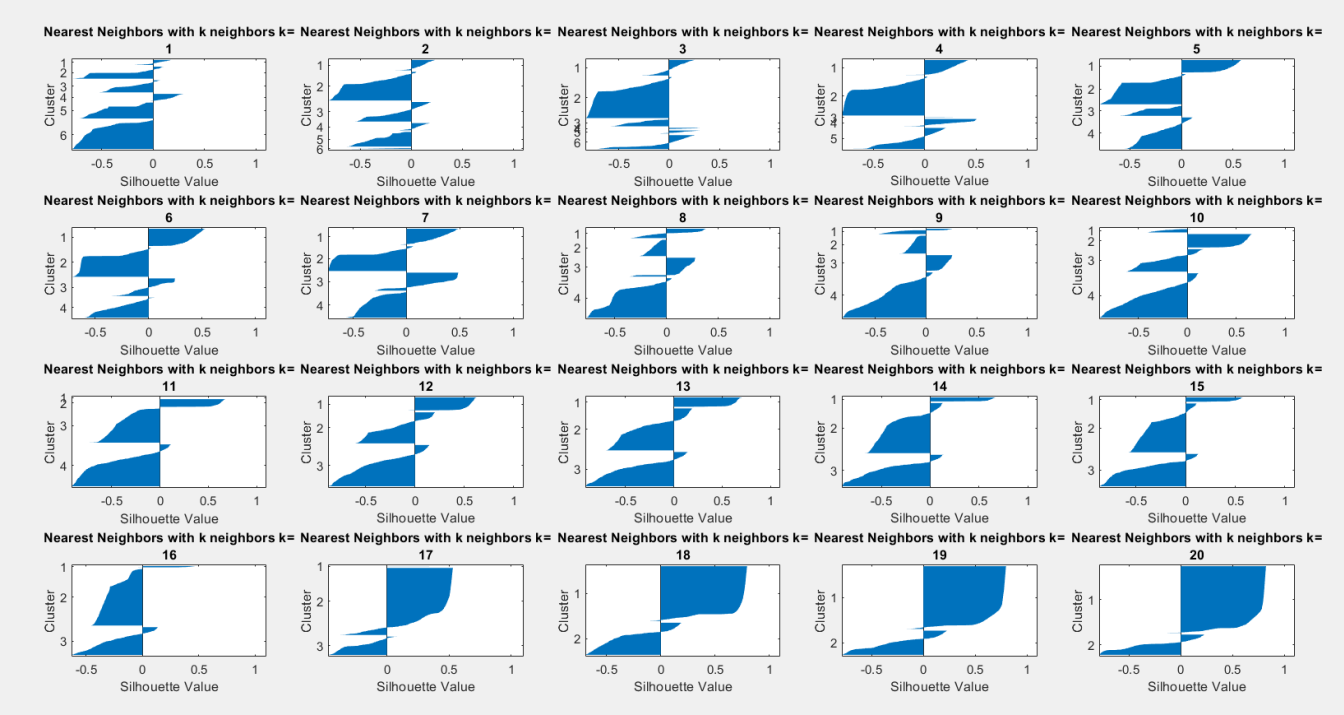


Figure 5.53 Charts of the average Silhouette value vs different values of the input parameter (K-Nearest neighbors). K-NEAREST NEIGHBORH method is considered.

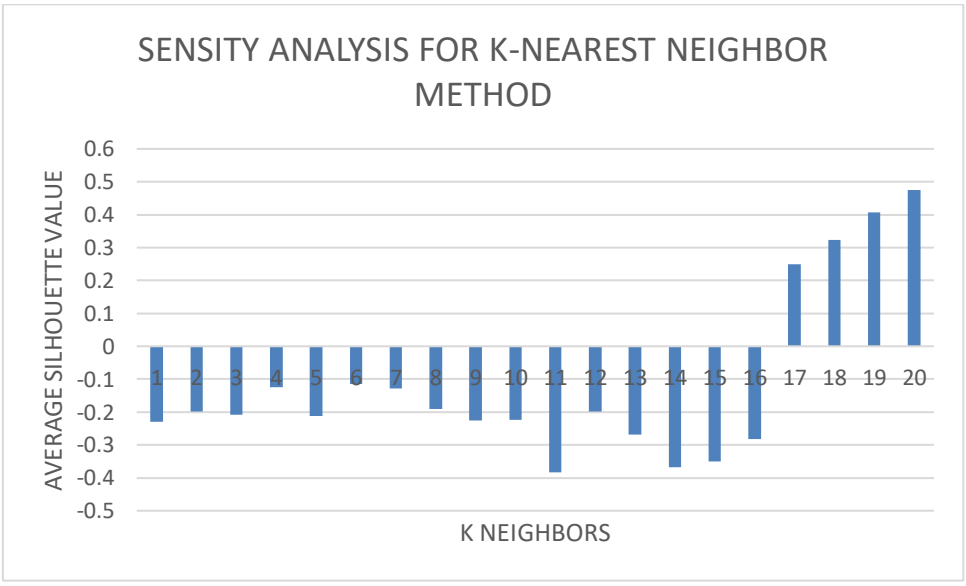


Figure 5.54 Chart of the average Silhouette values different values of the input parameter (K-Nearest neighbors). K-NEAREST NEIGHBOR method is considered.

It is seen that the silhouette values for different values of input parameters do not exceed 0.50. Which gives us an idea, that its performance of the K-Nearest Neighbors is low. Additionally, there are some negative Silhouette values in almost all the cases in the figure 5.50.

KMEAN METHOD

FROM CLUSTER K=1 TO CLUSTER K=20

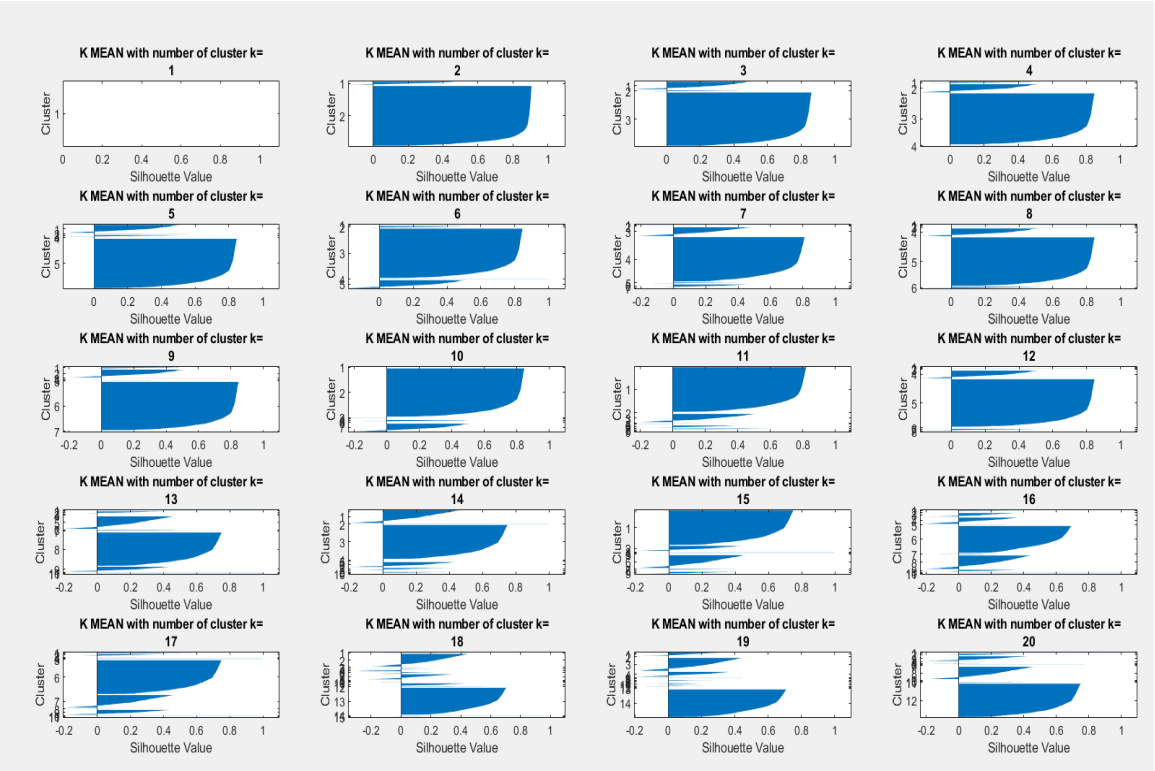


Figure 5.55 Charts of the average Silhouette value vs different values of the input parameter (K-clusters). KMEAN method is considered.

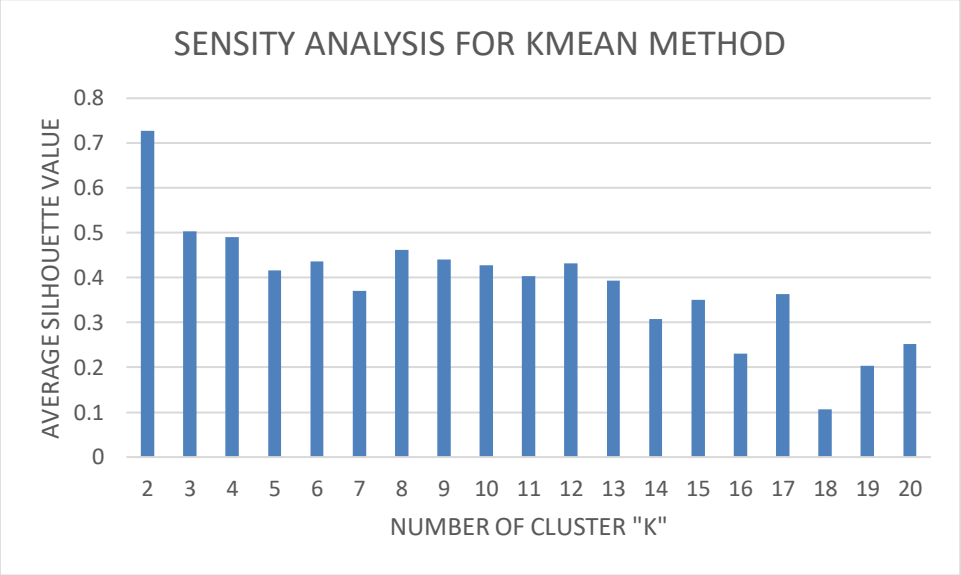


Figure 5.56 Chart of the average Silhouette values different values of the input parameter (K-clusters). KMEAN method is considered.

The maximum silhouette value for the KMEAN method is when the cluster number K is equal to 2. Additionally, as the number of clusters increases, the silhouette value of each group decreases. An additionally fact is when K= 1, there is not solution, it happens because the KMEAN method needs at least 2 clusters to be run.

5.7 Elbow Chart Analysis

In this section it will be seen the graphs of the Elbow method using the distortion criterion and the inertia criterion for each method analyzed. Recalling, the distortion is a measure that gives how far apart the groups are. The larger it is, the less similar are the objects being compared. Generally, the bigger the better. On the other hand, inertia tells me how similar the members are in a group. The smaller the value, the better.

The best k (number of clusters) is assumed to be in which the Elbow Method graph has an inflection (an elbow). The Elbow chart is more used to observe the adequate number of clusters for KMEAN methods.

DBSCAN METHOD

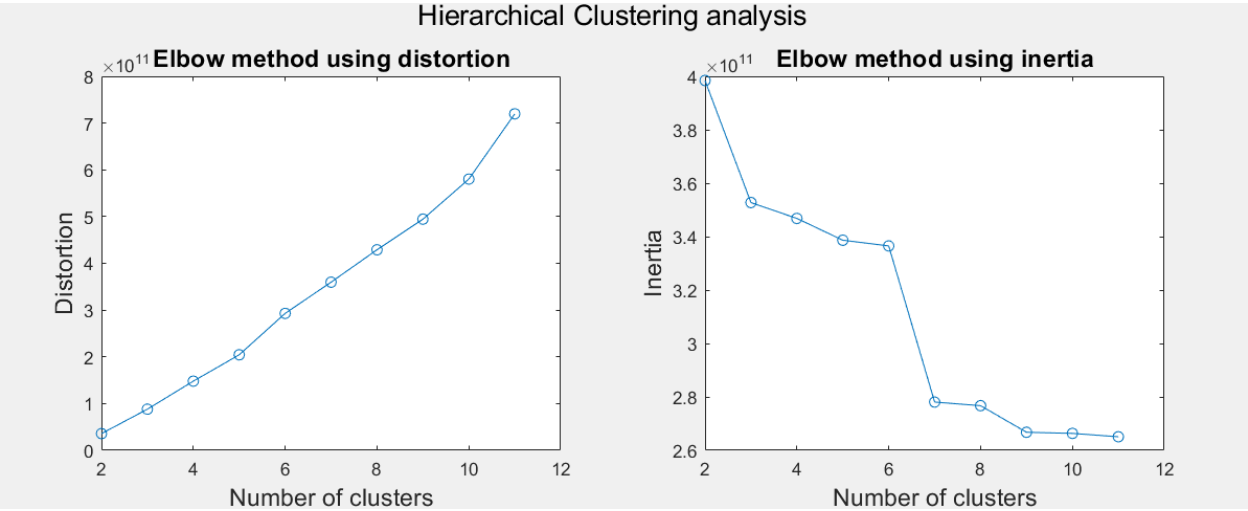


Figure 5.57 Elbow charts using distortion and inertia method. DBSCAN clustering is considered.

According to the above charts, in the DBSCAN method neither the distortion criterion nor the inertia criterion gives a clear elbow (there are some of them).

HIERARCHICAL METHOD

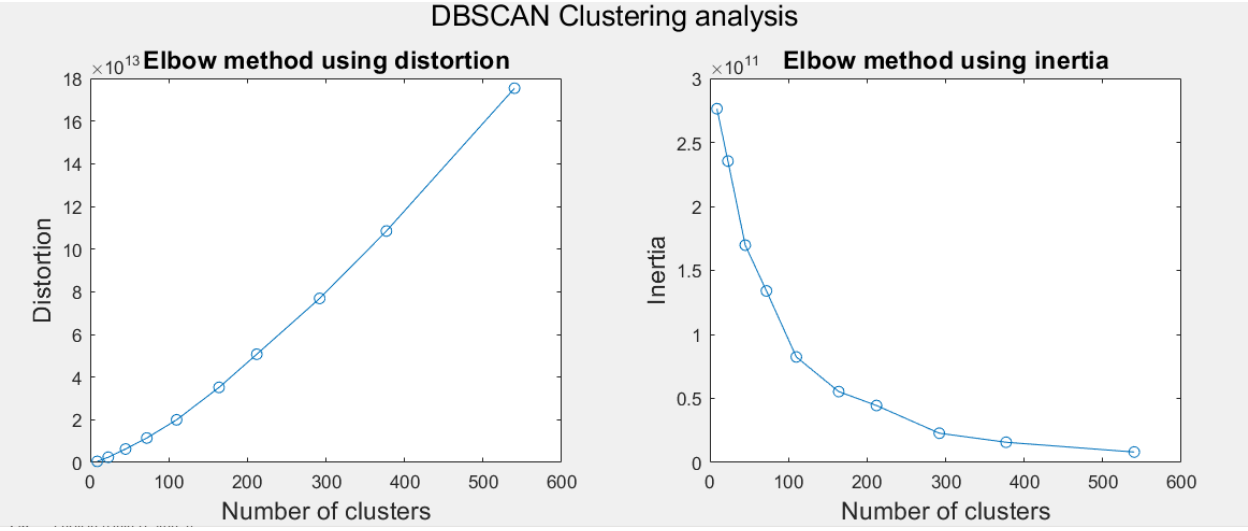


Figure 5.58 Elbow charts using distortion and inertia method. Hierarchical clustering is considered.

In the Hierarchical method, the same happened as the DBSCAN, there is no clear elbow neither the distortion criterion nor the inertia criterion. Both graphs are similar to a quadratic function with no change in slope.

K-NEAREST NEIGHBOR METHOD

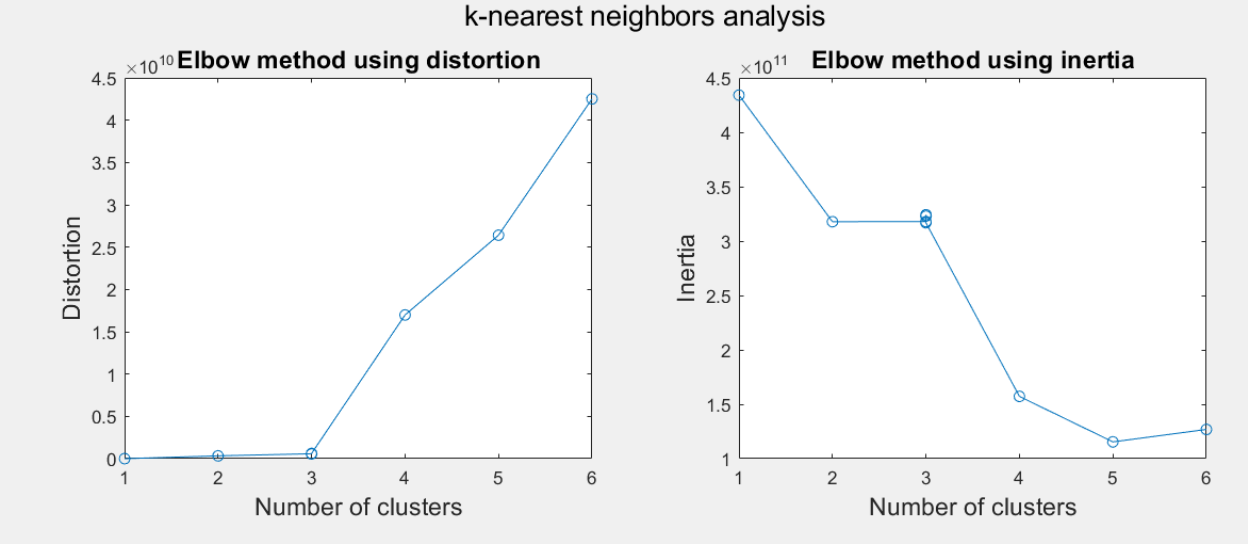


Figure 5.59 Elbow charts using distortion and inertia method. K-Nearest Neighbors clustering is considered.

In the K-Nearest Neighbor method, the Elbow graph neither using the distortion nor the inertia criterion, there is an expected smooth graph form, there are many elbows.

K- MEAN METHOD

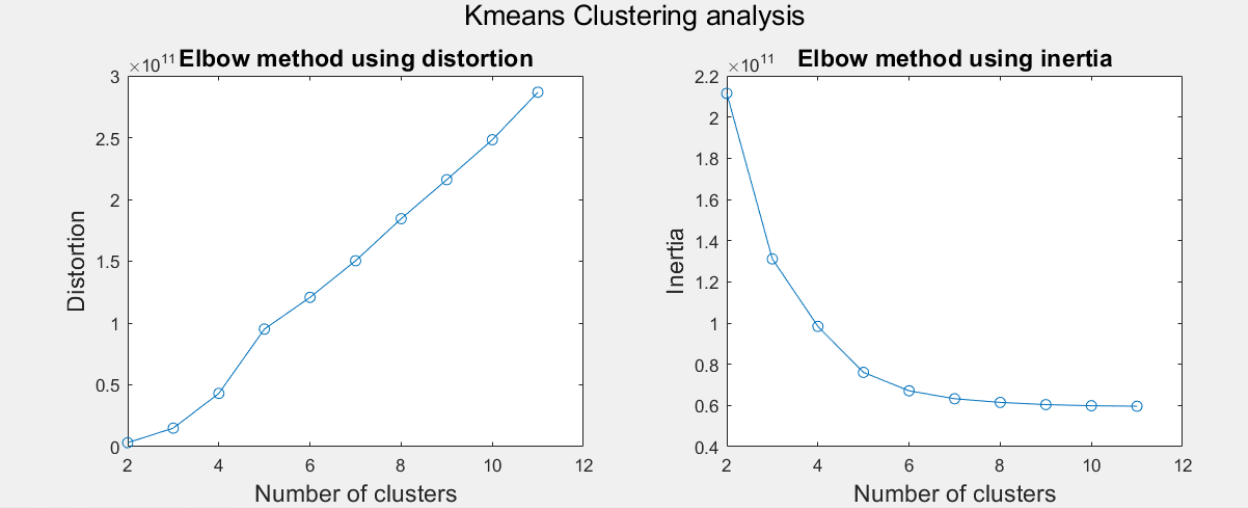


Figure 5.60 Elbow charts using distortion and inertia method. K-Mean clustering is considered

.

For the case of the KMEAN method, the Elbow plot using both criteria tell that there is an elbow for “K = 5 clusters”. Based on this, it can be said that according to this graph, the appropriate number of clusters is when 5 clusters are considered.

5.8 Silhouette charts and data dispersion considering the whole database

In this section each method will be discussed using the silhouette chart. Along with this graph, a dispersion graph of each method is presented in order to observe how the data are located in a chart made based on the PCA scores.

The silhouette chart of each method was made using the input parameters required by each method based on the results of the sensitivity analysis of each method for the entire dataset.

DBSCAN METHOD

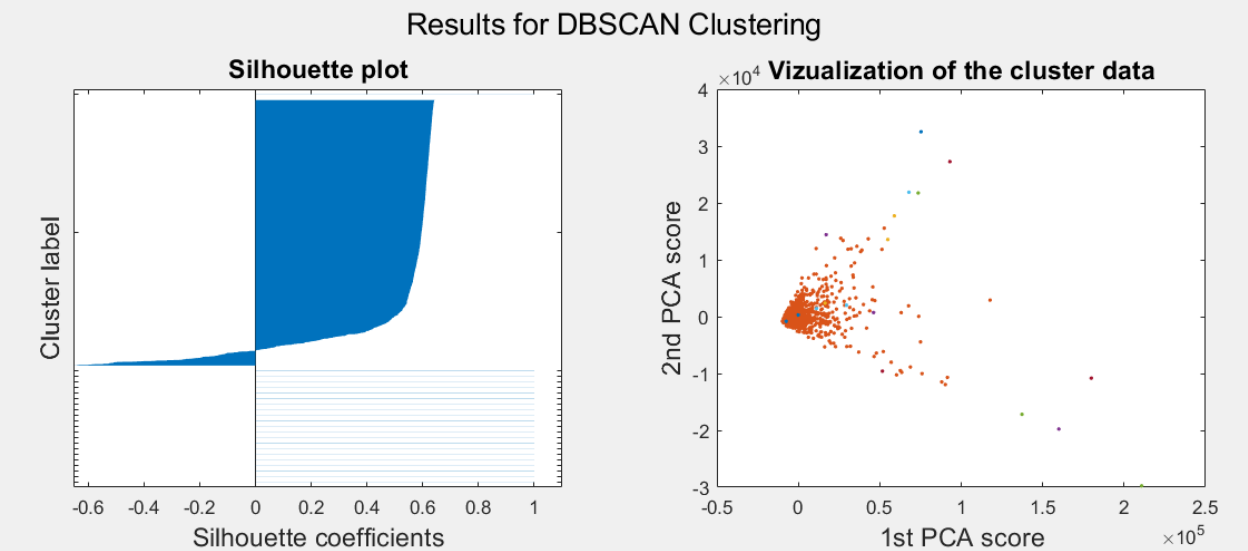


Figure 5.61 Charts of the average Silhouette value and data dispersion for DBSCAN method. The entire data is considered.

HIERARCHICAL METHOD

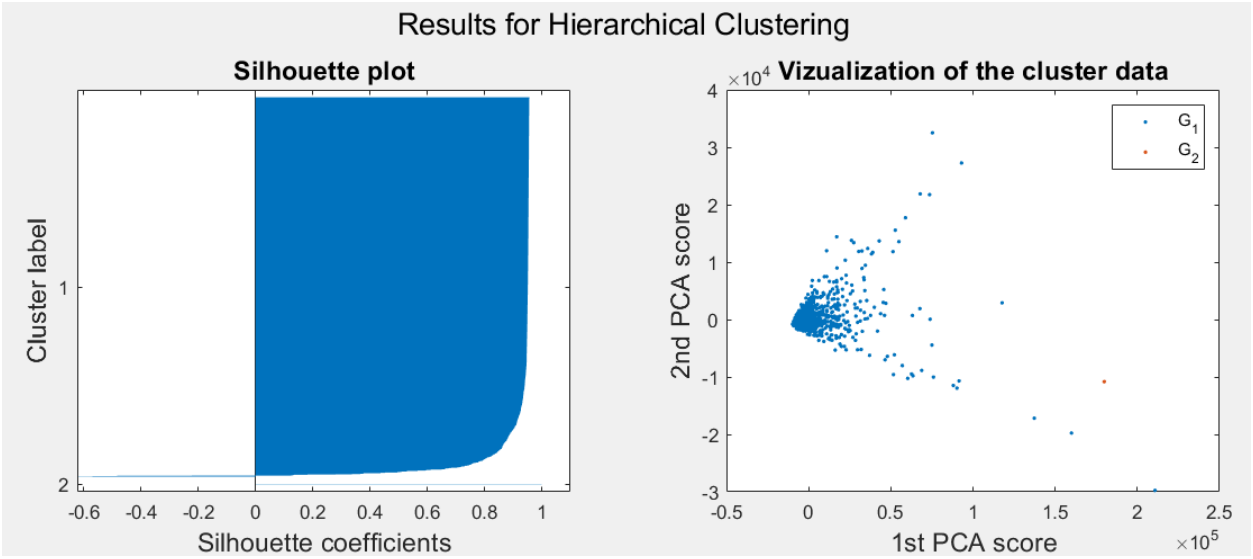


Figure 5.62 Charts of the average Silhouette value and data dispersion for Hierarchical method. The entire data is considered.

K- NEAREST NEIGHBOR METHOD

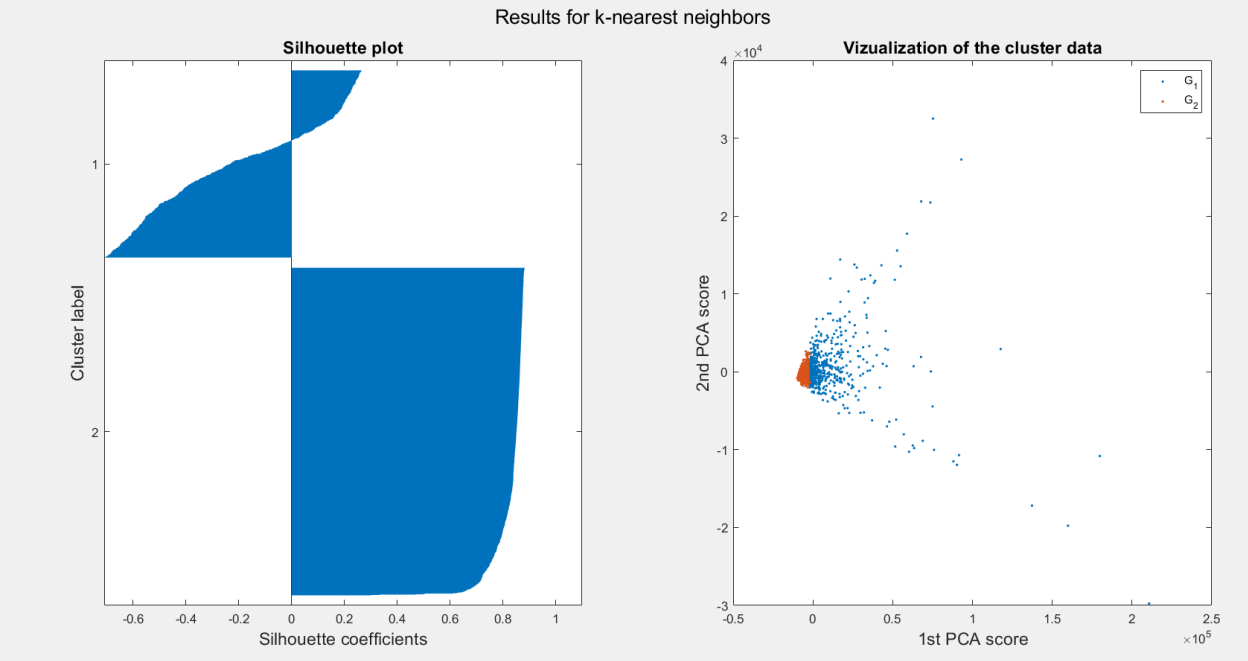


Figure 5.63 Charts of the average Silhouette value and data dispersion for K-Nearest Neighbors method. The entire data is considered.

K-MEAN METHOD, K=2 CLUSTERS

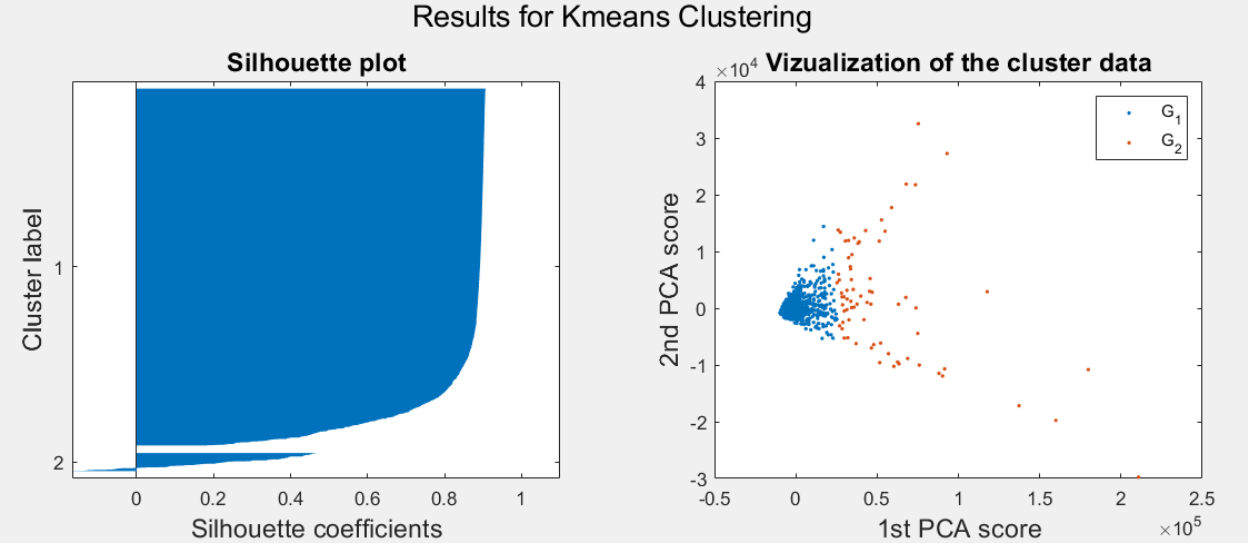


Figure 5.64 Charts of the average Silhouette value and data dispersion for K-Mean method with K=2 clusters. The entire data is considered.

K-MEAN METHOD, K=3 CLUSTERS

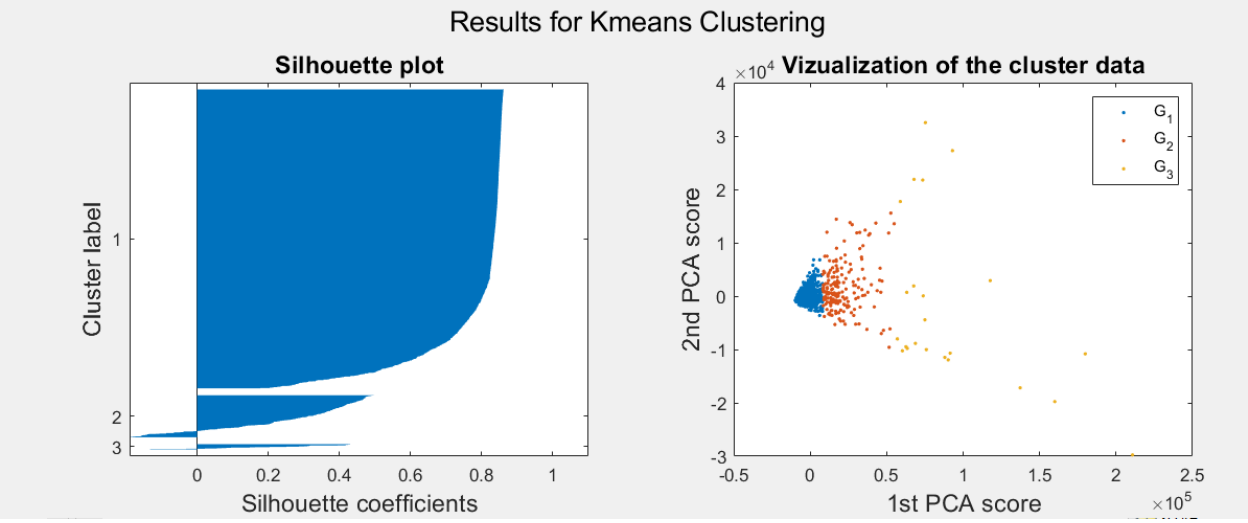


Figure 5.65 Charts of the average Silhouette value and data dispersion for K-Mean method with K=3 clusters. The entire data is considered.

K-MEAN METHOD, K=4 CLUSTERS

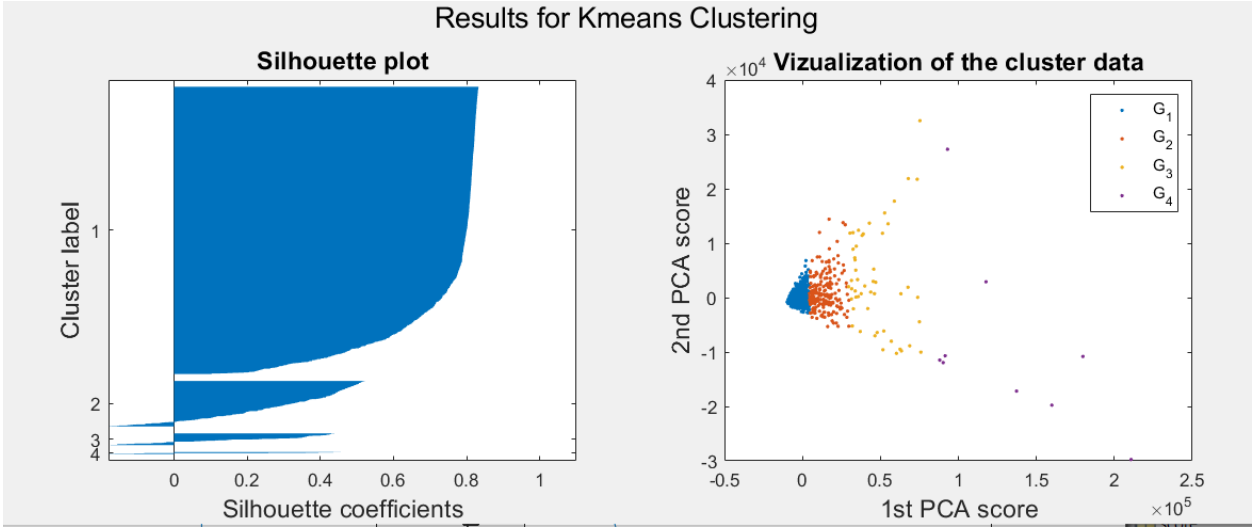


Figure 5.66 Charts of the average Silhouette value and data dispersion for K-Mean method with K=4 clusters. The entire data is considered

.

K-MEAN METHOD, K=5 CLUSTERS

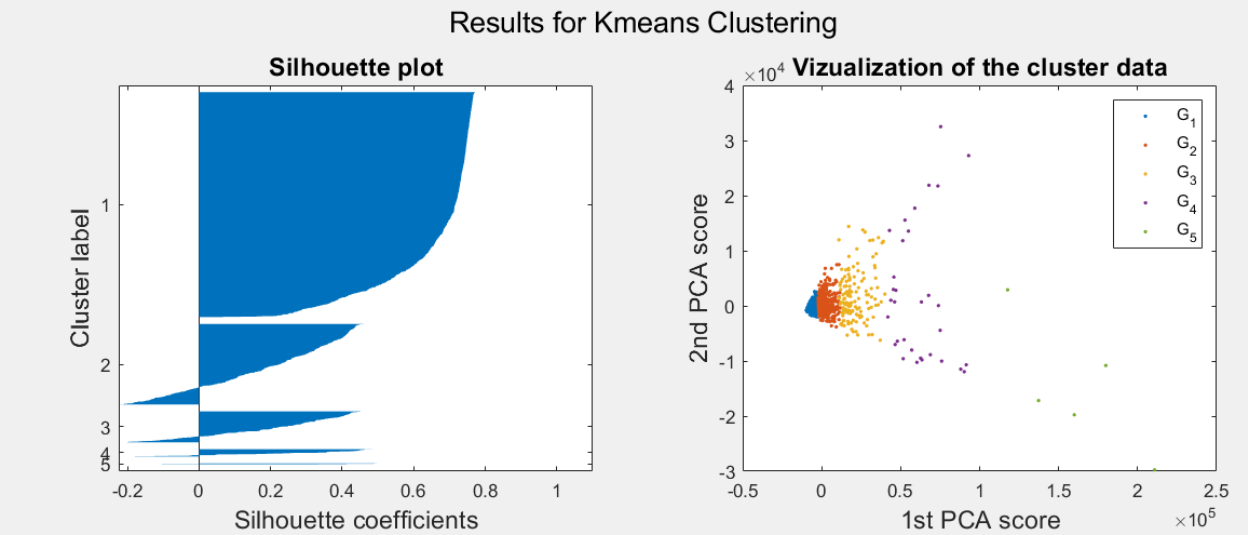


Figure 5.67 Charts of the average Silhouette value and data dispersion for K- Mean method with K= 5 clusters. The entire data is considered.

5.9 Normalization of the curves for each cluster

This section shows each method analyzed with the silhouette graph and the normalized curves for each cluster, which were calculated with the mean of each power curve of each POD of the group.

Graphs are also showing the power used as a reference value for the normalization of the curve of each group and for each method.

DBSCAN METHOD

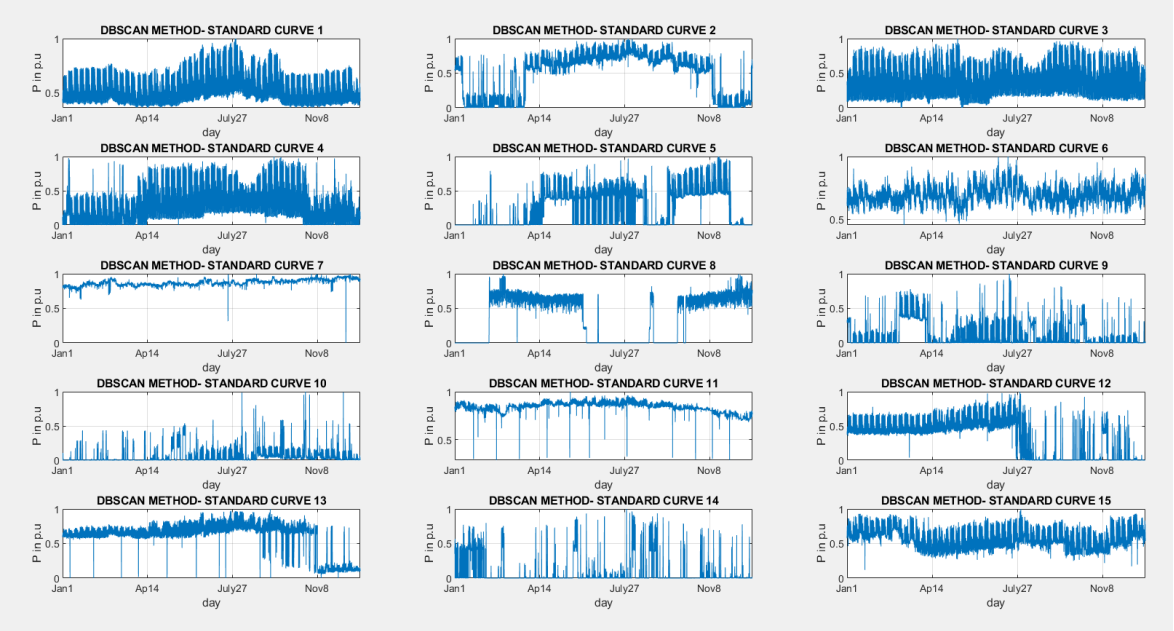


Figure 5.68 Normalized Power Curves for each cluster made using the DBSCAN method.

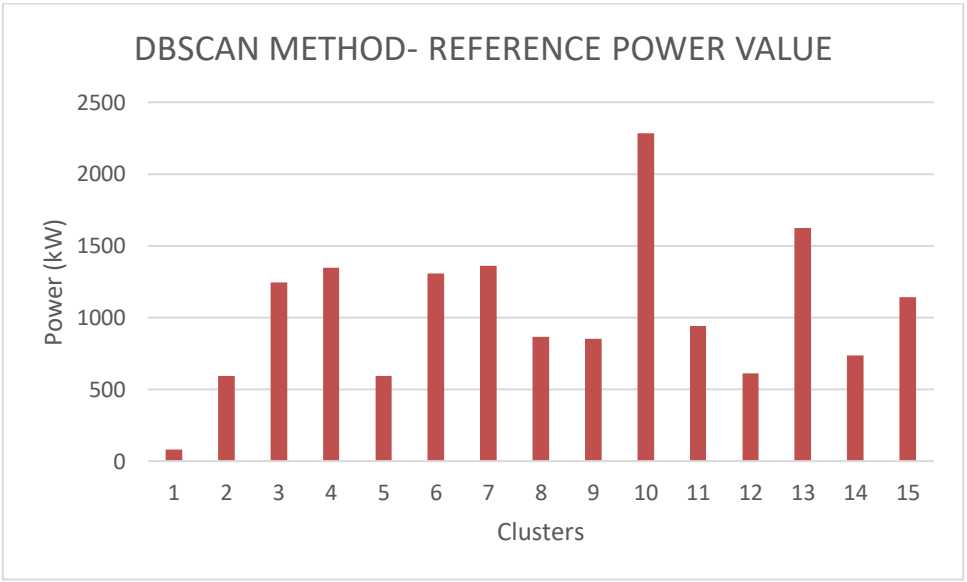


Figure 5.69 Device PODs Power Reference Value for each cluster. DBSCAN method is considered.

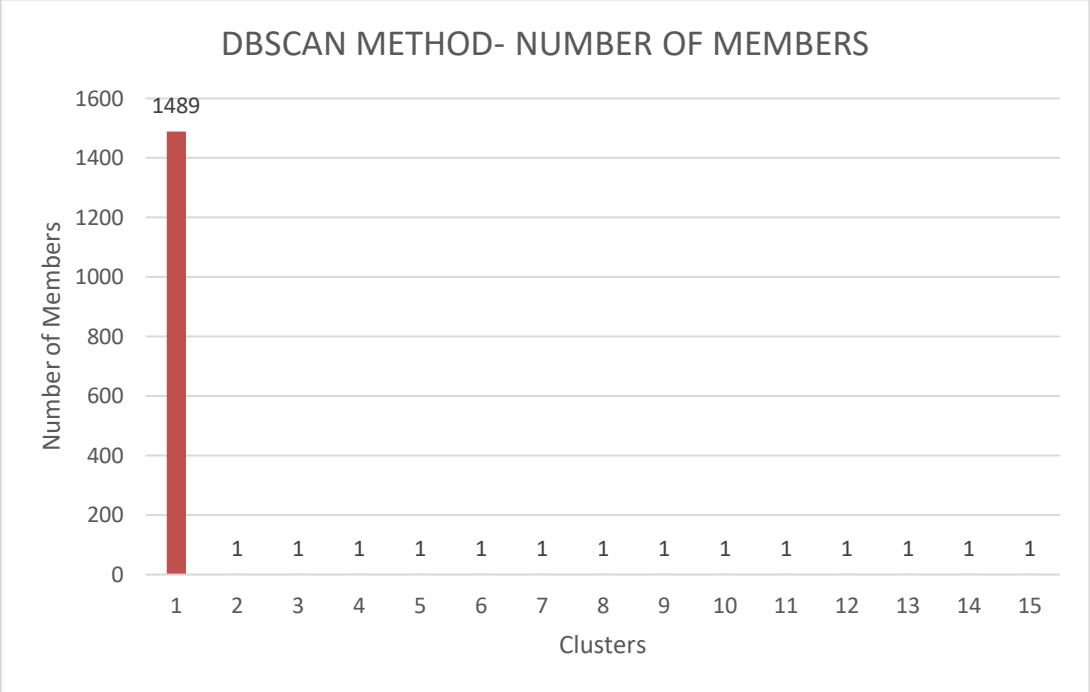


Figure 5.70 Number of members for each cluster. DBSCAN method is considered.

HIERARCHICAL METHOD

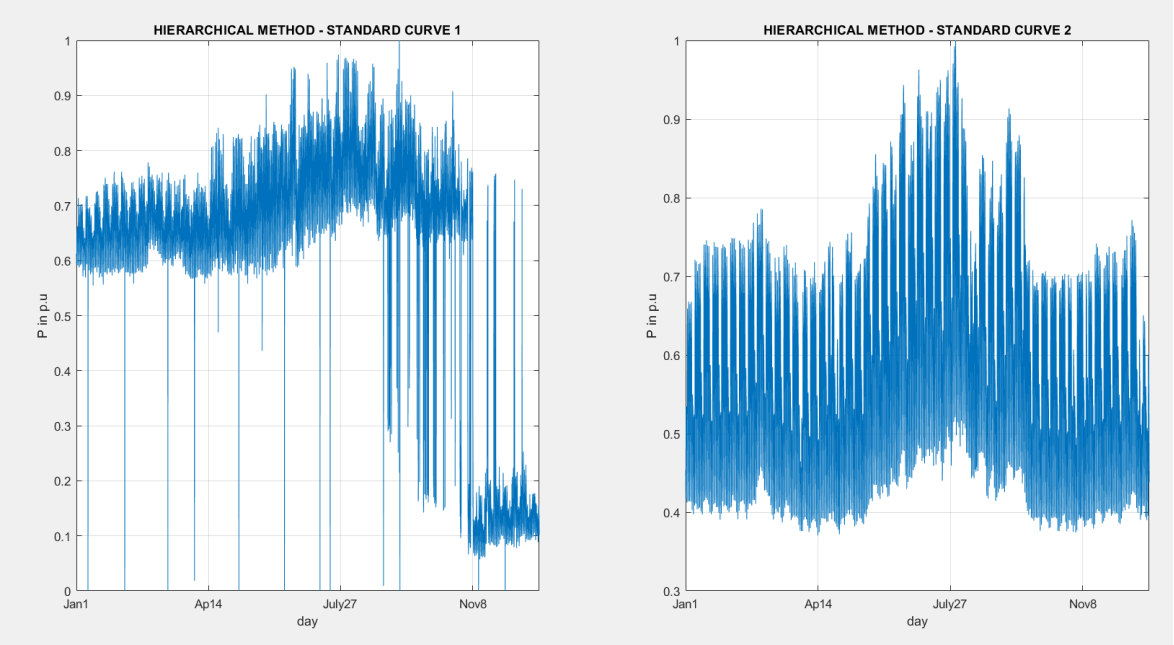


Figure 5.71 Normalized Power Curves for each cluster made using the Hierarchical method.

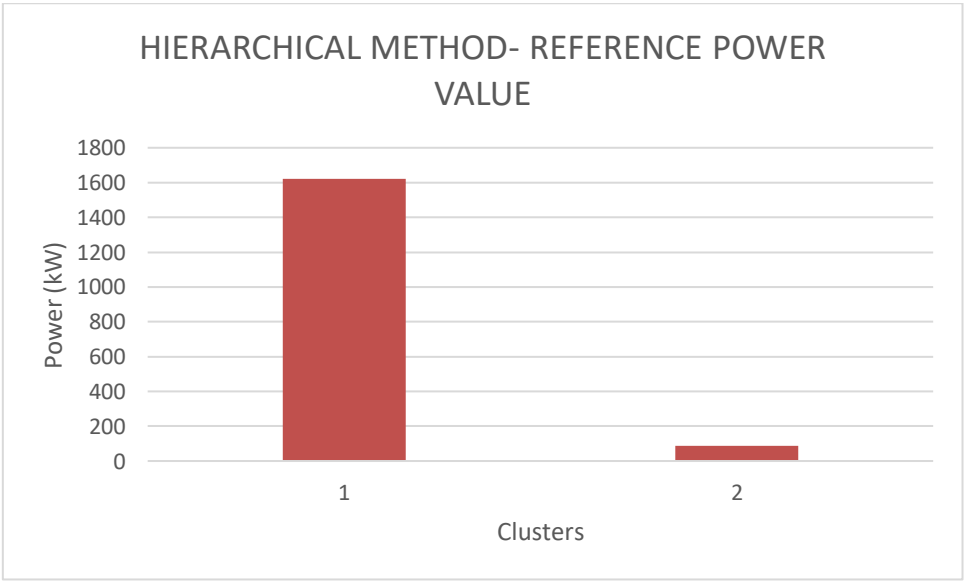


Figure 5.72 Device PODs Power Reference Value for each cluster. Hierarchical method is considered.

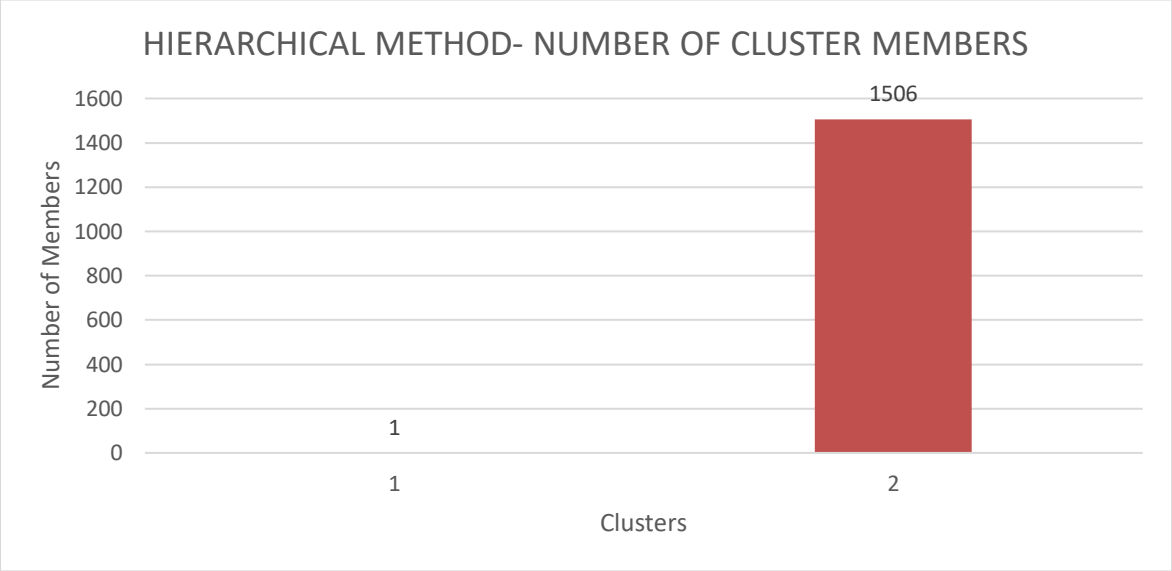


Figure 5.73 Number of members for each cluster. Hierarchical method is considered.

K-NEAREST NEIGHBORS METHOD

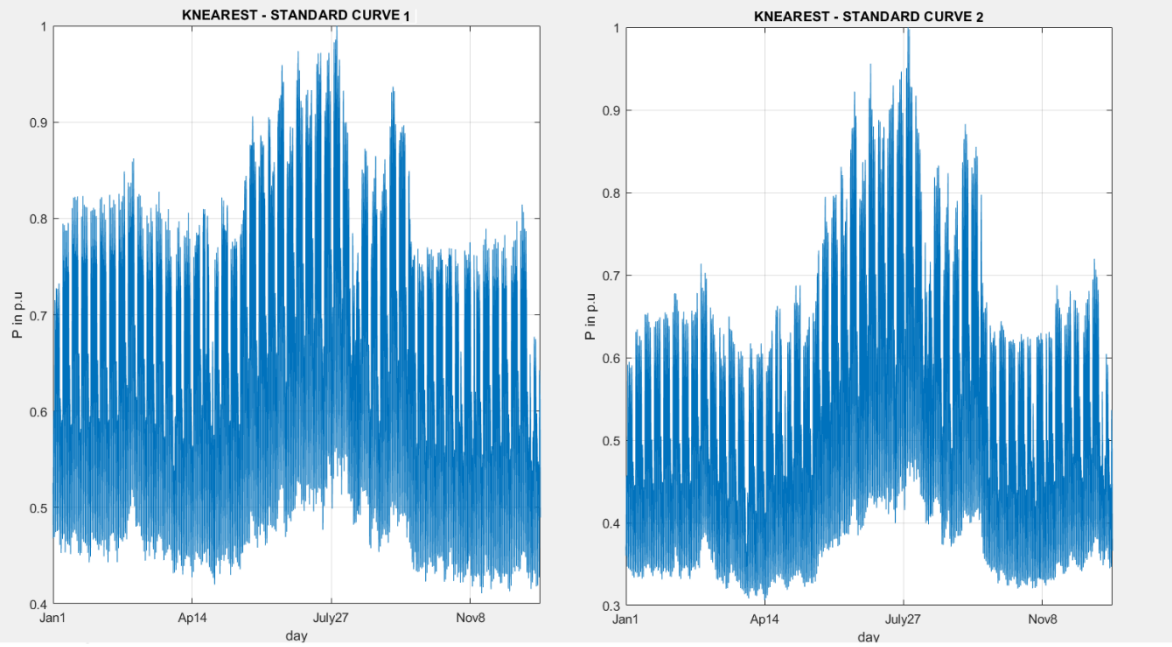


Figure 5.74 Normalized Power Curves for each cluster made using the K-Nearest Neighbor method.

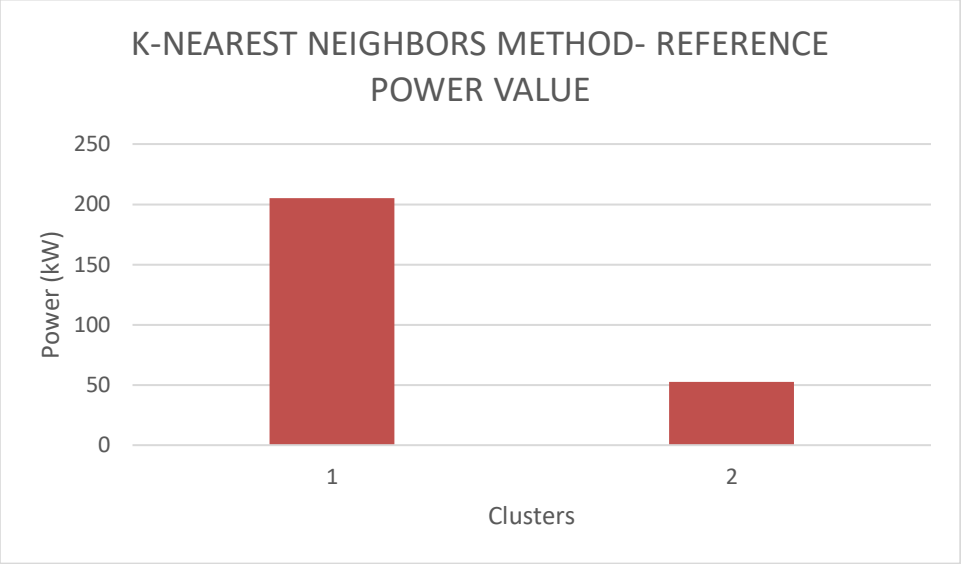


Figure 5.75 Device PODs Power Reference value for each cluster. K-Nearest Neighbor method is considered.

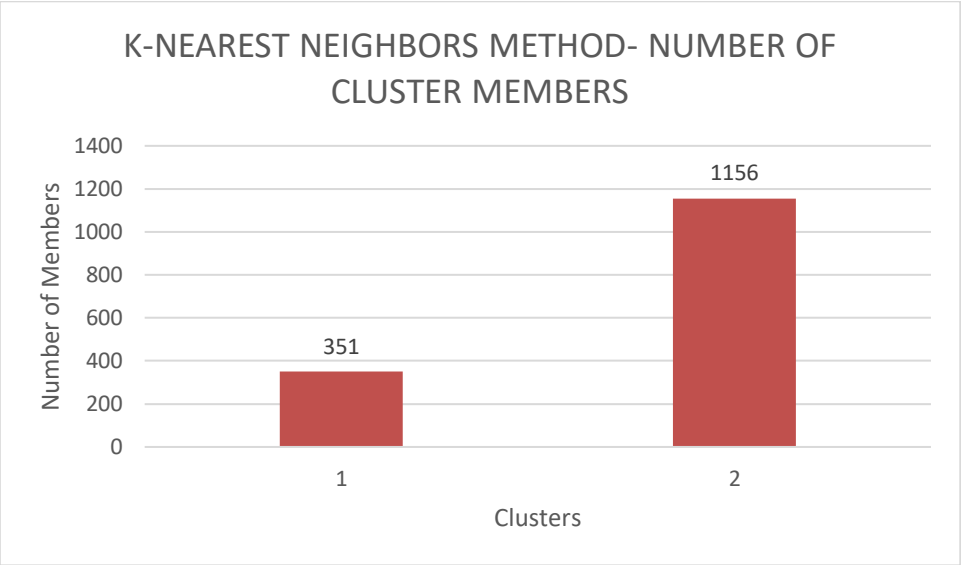


Figure 5.76 Number of members for each cluster. K-Nearest Neighbor method is considered.

KMEAN METHOD CONSIDERING K=2 CLUSTERS

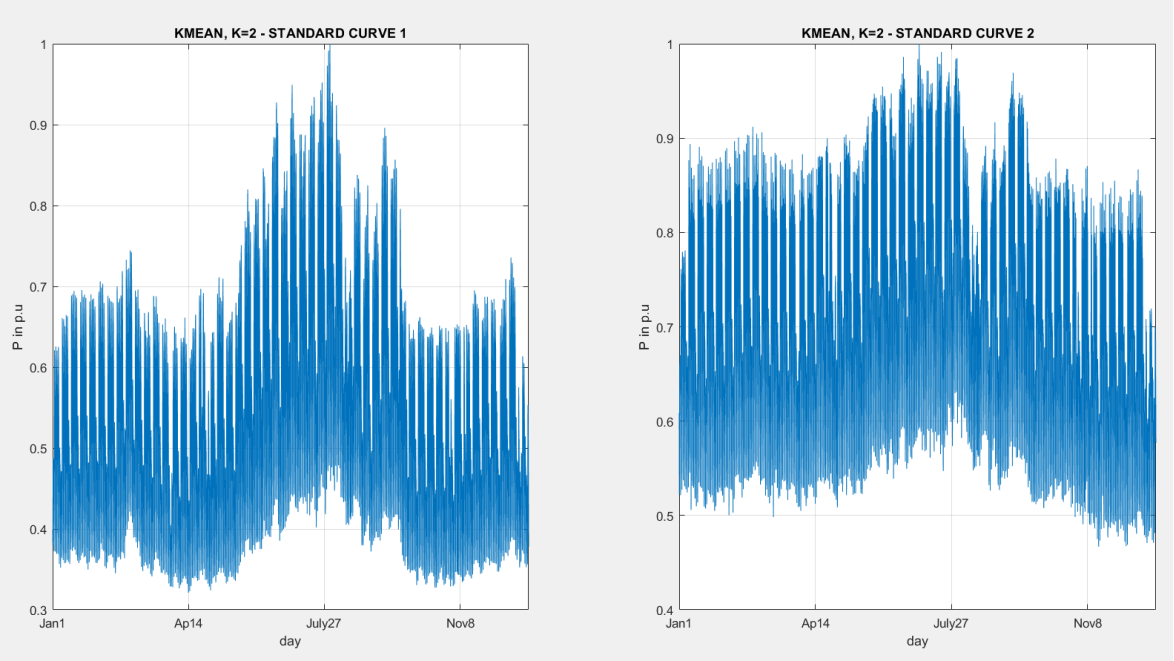


Figure 5.77 Normalized Power Curves for each cluster made using the K-Mean method with K= 2 cluster.

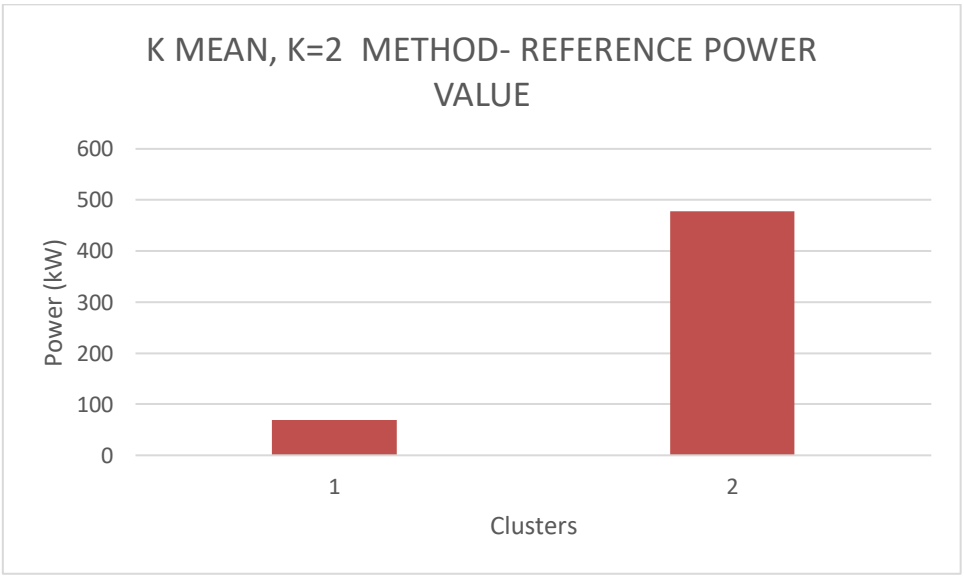


Figure 5.78 Device PODs Power Reference for each cluster. K-Mean method with k=2 clusters is considered.

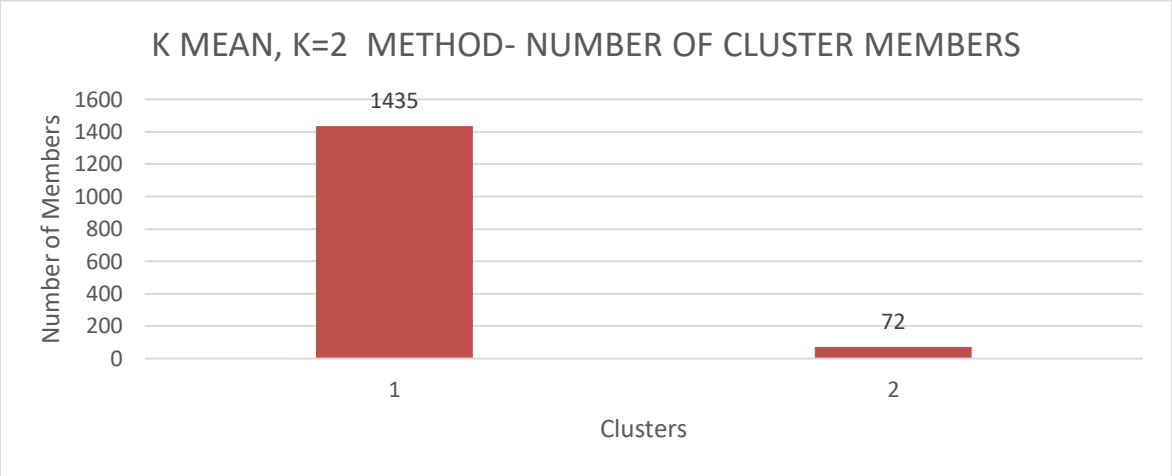


Figure 5.79 Number of members for each cluster. K-Mean (K=2 clusters) method is considered.

KMEAN METHOD CONSIDERING K=3 CLUSTERS

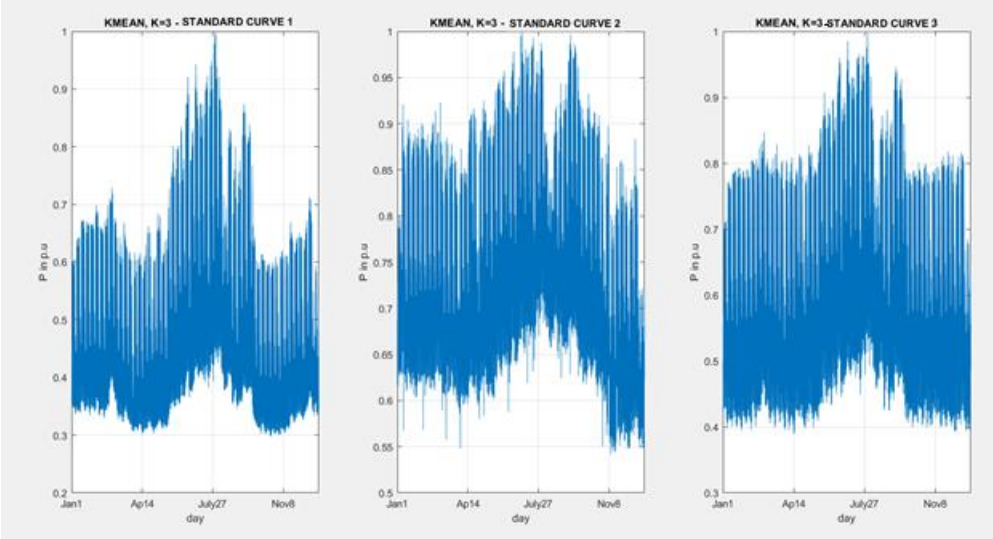


Figure 5.80 Normalized Power Curves for each cluster made using the K-Mean method with K= 3 cluster.

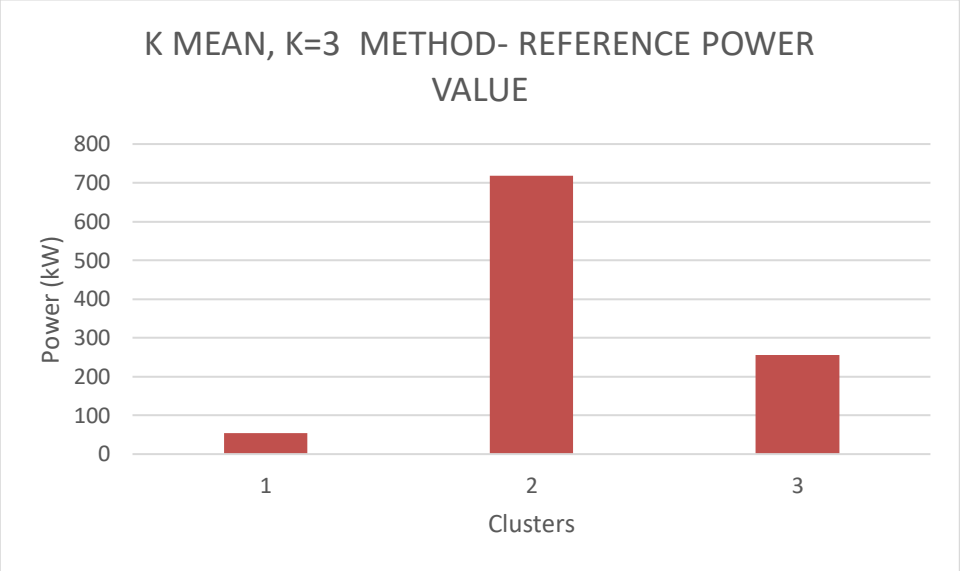


Figure 5.81 Device PODs Power Reference Value for each cluster. K-Mean method with k=3 clusters is considered.

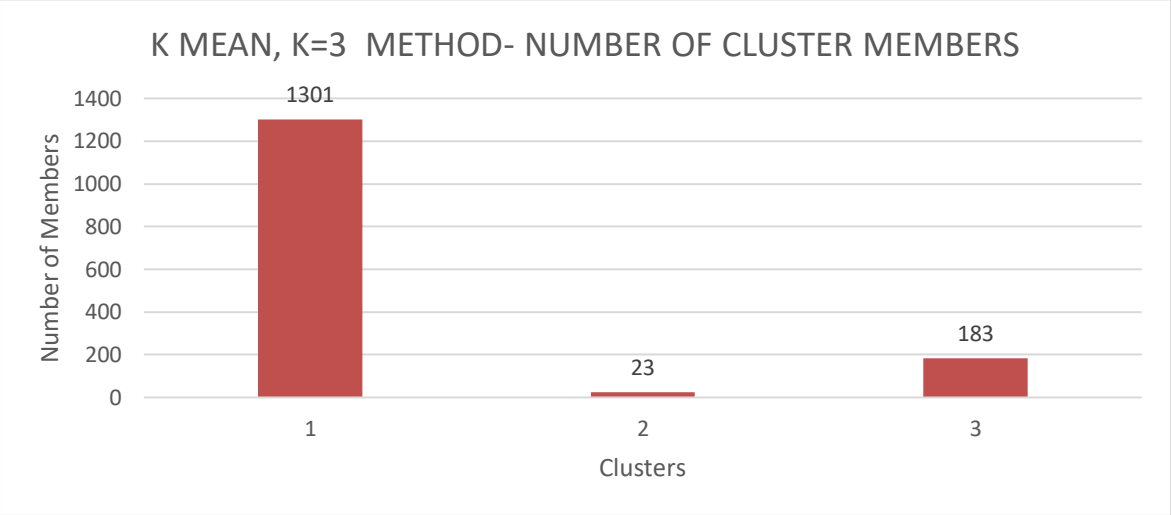


Figure 5.82 Number of members for each cluster. K-Mean (K=3 clusters) method is considered.

KMEAN METHOD CONSIDERING K=4 CLUSTERS

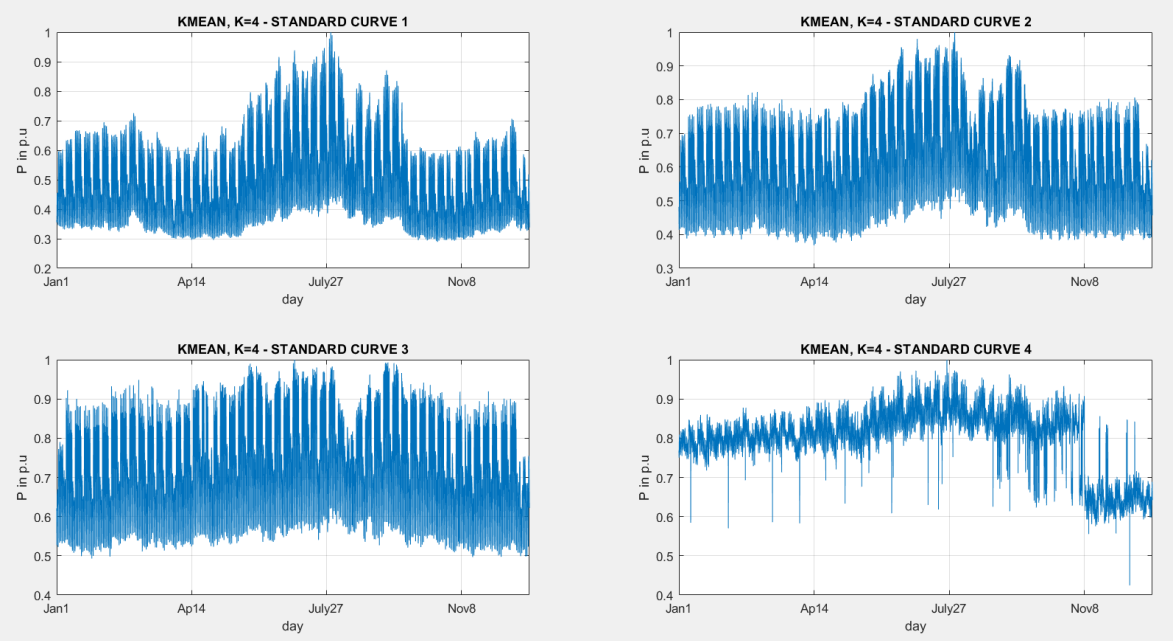


Figure 5.83 Normalized Power Curves for each cluster made using the K-Mean method with K= 4 cluster.

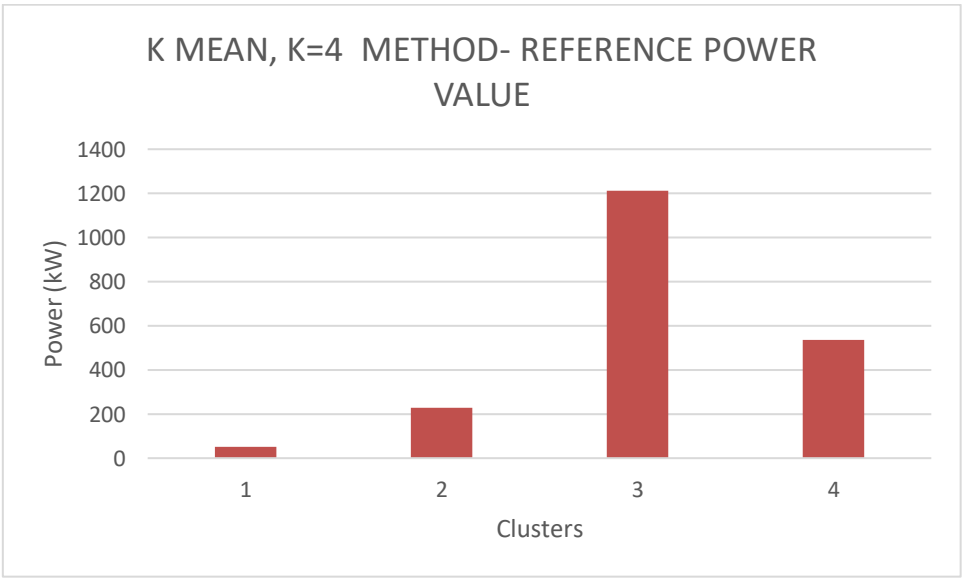


Figure 5.84 Device PODs Power Reference value for each cluster. K-Mean method with k=4 clusters is considered.

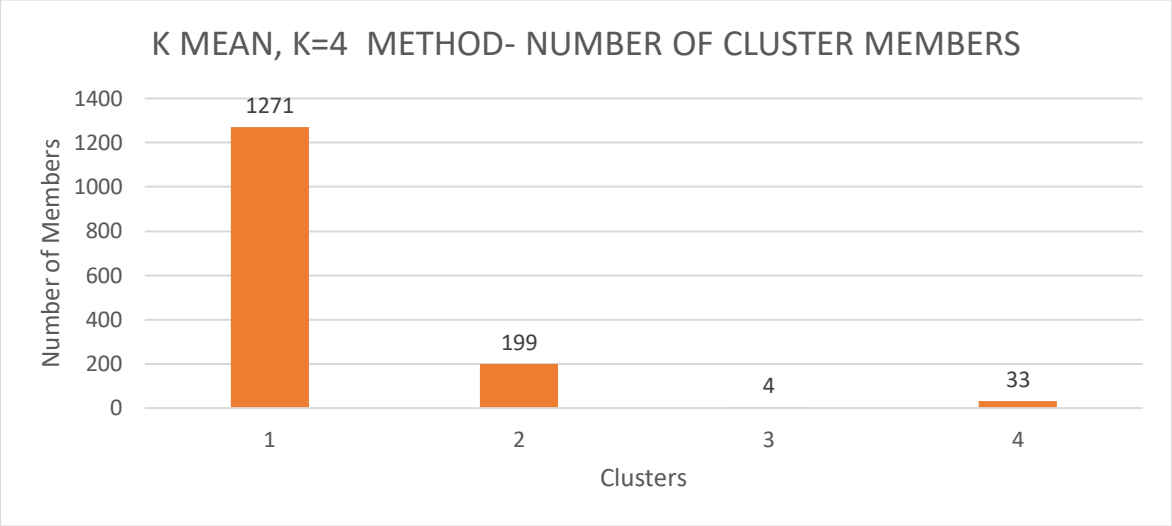


Figure 5.85 Number of members for each cluster. K-Mean (K=4 clusters) method is considered.

KMEAN METHOD CONSIDERING K=5 CLUSTERS

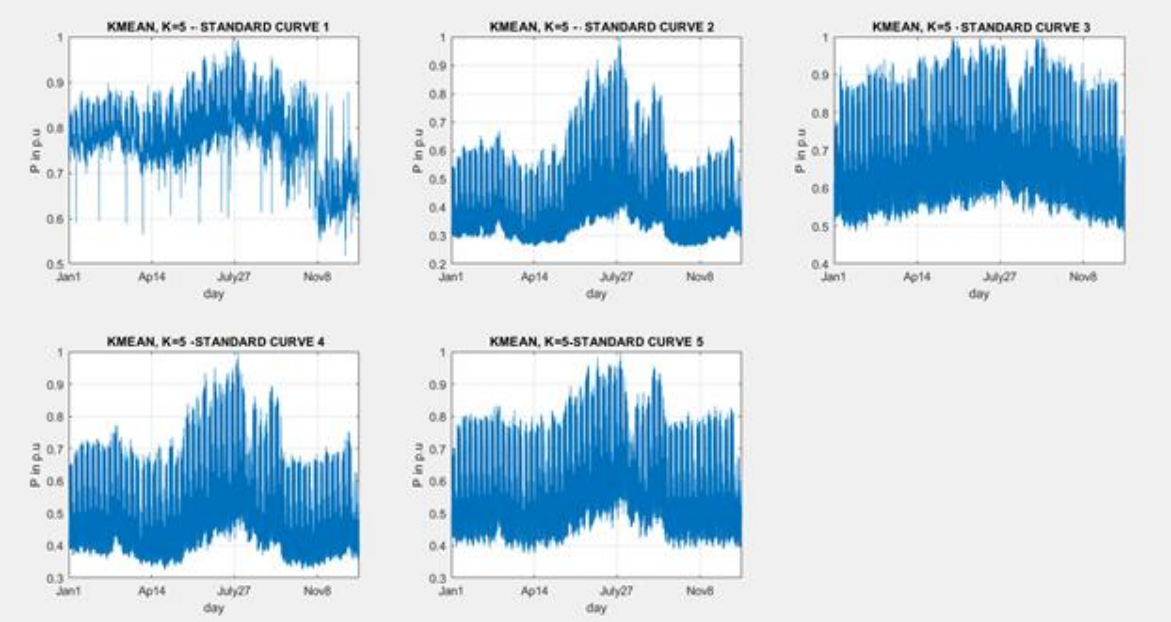


Figure 5.86 Normalized Power Curves for each cluster made using the K-Mean method with K= 5 cluster.

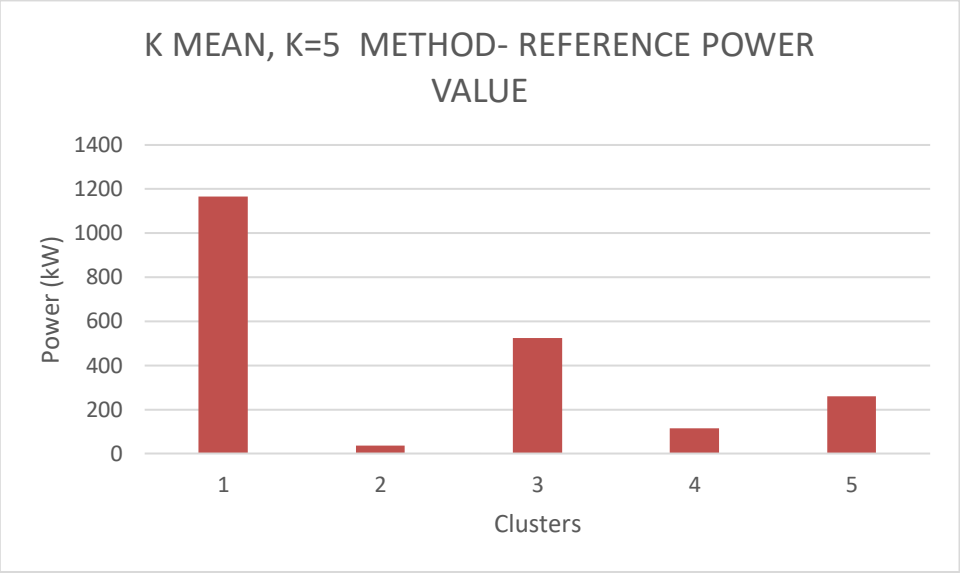


Figure 5.87 Device PODs Power Reference Value for each cluster. K-Mean method with k=5 clusters is considered.

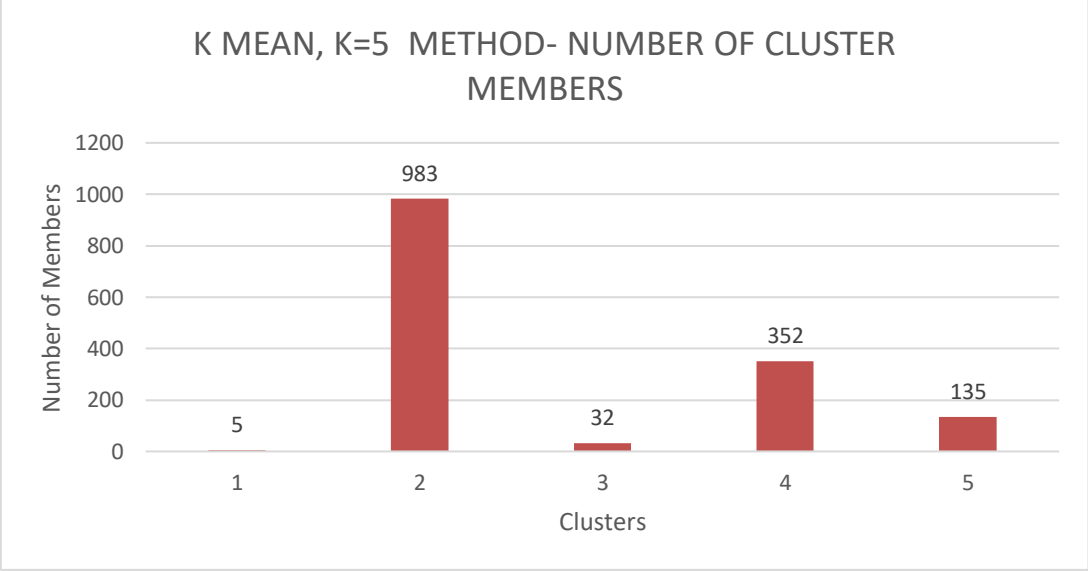


Figure 5.88 Number of members for each cluster. K-Mean (K=5 clusters) method is considered.

From the previous graphs, it can be seen that there is a trend for large groups of having a reference value (which is the average of all the powers of each group's POD) less than 100 kW, while groups of few elements, this reference value is bigger than 1000 kW. This behavior is excepted to single-member groups, as is the case with the K-Nearest Neighbor method

The KMEAN method with K = 2 clusters, gives an average silhouette value close to 0.75. It can be seen that one of the two groups have a large number of grouped measurement devices, 1435 PODs. While the smallest group has the rest of data, about 72 PODs.

Additionally, the normalized power curve obtained for each cluster is very important for the planning offices of the electricity companies.

5.10 Conclusions of the results section

A sensitivity analysis was performed for each clustering method, using three different and random sample sets. This work was done based on the input parameters, using a MATLAB code. Many combinations of these parameters were put into evaluation, and the best combination was selected. Then, with the appropriate arrangement that guarantees the maximization of the silhouette value for each method considering every representation and every time-step, it proceeded to look for the best representation of data and time-step. It was found that the best scores were obtained when the absolute value with a time-step of 15 minutes was considered.

Once this was found, a process was carried out for seeing which the best method. For the samples, the Hierarchical and KMEAN methods, with K = 2 cluster gave the best outcomes based on the average silhouette value. With this idea in mind, the analysis was applied to the entire database.

Based on the analysis done on the samples, it was decided to use the combination of absolute values and a time step of 15 minutes. From there, the sensitivity analysis was carried out, varying, as was done with the samples, the input parameters to find the maximization of the silhouette values. As expected, the same thing happened with entire

dataset than the samples. In other words, the methods with the best outcomes were Hierarchical and KMEAN with k = 2 clusters (based on the average silhouette value).

These methods give us that all the data is divided into two groups, but in the case of Hierarchical one of the groups has only one member, so it is discarded as the best method. The next best performing method is KMEAN with k = 2 clusters. In this method it is observed that its two groups have many elements. It also is observed that one of the groups of this KMEAN, has most of the data, while the other group, the rest of data.

An additional analysis was carried out, the Elbow Chart in order to see the appropriated quantity of group. It was applied for all the methods in analysis, but it was seen (like the theory says) this approach is most valid for K-Mean methods. According to this, the appropriated number of clusters to be considered was 5, i.e. K-Mean with K=5 clusters.

6 CONCLUSIONS

This thesis mainly examined about a methodology to find the performance for each clustering method based on the silhouette values. An additional performance analysis using the Elbow charts were made. For each method in study, the normalized electric load curves of every group were obtained. This information is important for planning purposes. The methodology was made with the real load data from the UNARETI company in Milan.

The methodology includes the following steps:

a. The methodology includes the Charging load data from the raw database. As the database was taken from measurement devices, this information had to be processed using Microsoft Access and MATLAB.
b. Processed load data. As the database contains information on substations outside of Milan, and since the data desired is that only from the metropolitan area of Milan, it had to be previously filtered. Additionally, those devices that have blank values were not considered.
c. Simplification of the database, using the PCA method. The main components PCA were extracted, in order to decrease the dimension of the data. An additional application of the principal components was to get the chart in order to see the distribution of the real data.
d. Each clustering method is based on different input parameters, then from the total database different sets of samples were taken to evaluate which input parameter arrays give the best outcomes. This process was replicated for the entire database. A sensitivity analysis was used.
e. Once having the best amalgamation of input parameters for each method, best data representation and best time-step were obtained. From the study over the samples, the best results were obtained when the absolute values representation and 15 minutes time-step were selected. This process was carried out for the samples and then with the adequate scores it was replicated to the entire database.
f. Knowing that the best silhouette values are given when the absolute values and time-step of 15 minutes are considered, a sensitivity analysis was applied over the whole data in order to get the maximum silhouette values.

g. Based on what the sensitivity analysis of the entire dataset gives us, it was possible to compare all the clustering methods and see the performance of each them based on silhouette values, also discarding those single-member groups with a silhouette value equal to 1.
h. An additional analysis was carried out, the Elbow Chart in order to see the appropriated quantity of group. It was applied for all the methods in analysis, but it was seen (like the theory says) this approach is most valid for K-Mean methods.

The methodology had the following features:

a. The data processing was made by using real data collecting during one year in some substations of UNARETI in Milan. It means, this study case can be the basis for future developments.
b. The data collecting was made every 15 minutes. Based on this time step, different analysis with different time step were made in order to see if using different time frame, a better efficiency could be gotten. The original time-step (15 minutes) has the best performance because the accuracy of the measurements increases.
c. Three different representations were evaluated: absolute value representation and two different P.U. values with different power reference value. The absolute value representation showed the best performance.

By analyzing the load customers classifying in Milan, some conclusions could be drawn:

a. The best results were obtained when the analysis worked with absolute values representation and a time-step of 15 minutes. Initial tests were carried out on a set of randomly chosen samples, and according to the results, the best approach was applied for the entire database.
b. Cause of having a lot of information (more than 50 million data) and to reduce calculation times, it is advisable to use an additional method such as PCA, thus reducing the dimensionality of the matrix of the entire database and to ease the computations.
c. From the PCA diagram, where the first two principal components were considered, it can be seen that at least ninety percent of the data is represented in these components, for this reason in some of the most performance cases evaluated, there were a tendency of having two clusters.

d. Elbow chart gives the idea that 5 groups is the appropriate number of clusters to be considered. The analysis was applied to all the methods, being the K-Mean methods with the best results. The distortion and inertia principles were considered.
e. In the sensitivity analysis, those groups of a single member with silhouette scores equal to 1 were not considered in the average silhouette value, because they were considered as outliers.
f. Based on the average silhouette method, Hierarchical and KMEAN with 2 clusters were those with high score versus the rest of methods. In the case of Hierarchical one of the groups has only one member, so it is discarded as the best method. The next best performing method is KMEAN with k = 2 clusters. For the rest of the KMEAN methods, it means, KMEAN with K greater or equal than 3, there is also a tendency to have a large group with some of the PODs, and other small ones; something similar like the KMEAN with K = 2 clusters
g. Analyzing the normalized curve of the method with the best performance, that is, KMEAN with K = 2 clusters, it can be seen the following: the cluster 1, the big group, has peaks of power in the summer months (June, July, August) with a small decreasing in mid-August, The other group, the small one, has a slightly more regular behavior than the large group. There is also the peak of power in the summer months, but it is not so different from the load of the rest of the year.

7 REFERENCES

[1] A. Berizzi, A. Bosisio, T. Qi, “A Method to classify Substation Load Profiles Based on PCA”, Milan, MSc thesis, Politecnico di Milano, 2017

[2] B. K. Tripathy (VIT University, India), Hari Seetha (Vellore Institute of Technology – Andhra Pradesh, India) and M. N. Murty (IISC Bangalore, India), “Uncertainty-Based Clustering Algorithms for Large Data Sets”. [Online]. Available: https://www.igi-global.com.

[3] Mark S. Aldenderfer & Roger K. Blashfield, “Cluster Analysis”, United States: Sage Publications, Inc, 1984

[4] R. Xu, D. Wunsch, “Clustering”, vol.1, John Wiley & Sons, Inc., Hoboken, New Jersey, 2009, pp. 1-12.

[5] M. Anderberg, “Cluster Analysis for applications”, Academic Press, Inc, New York, 1973, pp. 1-8.

[6] N. Jankowski, K. Grbczewsk, “Learning Machines”, Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on, 2007.

[7] T. Rashid, “Clustering”

http://www.cs.bris.ac.uk/home/tr1690/documentation/fuzzy_clustering_initial_report/node11.html

[8] J. Aronson, L. Iyer, “Cluster Analysis”, United States: SpringerLink, 2001 Edition.

[9] S. Kotsiantis, P. E. Pintelas, “Recent Advances in Clustering: A Brief Survey”, University of Patras, WSEAS Transactions on Information Science and Applications, 200

[10] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on*

Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297.

[11] L. Guiviant Viola, “Clustering electricity usage profiles with K-means”, *2018 Towards Data Science medium publication*. [Online]. Available: https://towardsdatascience.com/clustering-electricity-profiles-with-k-means.html.

[12] M. Pathak, “Hierarchical Clustering in R”, *2018 DataCamp*. [Online]. Available: https://www.datacamp.com/community/tutorials/hierarchical-clustering-R.html.

[13] MATLAB 2020a, The MathWorks Inc. [“Choose Cluster Analysis Method”]. Natick, Massachusetts, United States; [2020].

[14] MATLAB 2020a, The MathWorks Inc. [“K-Means Clustering”]. Natick, Massachusetts, United States; [2020].

[15] MATLAB 2020a, The MathWorks Inc. [“Hierarchical Clustering”]. Natick, Massachusetts, United State; [2020].

[16] T. Tullis, B. Albert, “Hierarchical Cluster Analysis”, Measuring the User experience (2nd Edition), Science Direct, 2013

[17] S. Sayad, “An introduction to Data Science- Clustering Method”, Master Program in Data Science of Rutgers University, New Jersey, 2010.

[18] MATLAB 2020a, The MathWorks Inc. [“DBSCAN”]. Natick, Massachusetts, United States; [2020].

[19] MATLAB 2020a, The MathWorks Inc. [“Cluster using Gaussian Mixture Model”]. Natick, Massachusetts, United States; [2020].

[20] MATLAB 2020a, The MathWorks Inc. [“Classification using Nearest Neighbors”]. Natick, Massachusetts, United State; [2020].

[21] Ester, M., H.-P. Kriegel, J. Sander, and X. Xiaowei. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data Mining*, 226-231. Portland, OR: AAAI Press, 1996.

[22] MATLAB 2020a, The MathWorks Inc. ["Comparison of Clustering Methods"]. Natick, Massachusetts, United States; [2020].

[23] L. Peterson (2009), *"K-nearest ne*ighbor", Center for Biostatistics, The Methodist Hospital Research Institute, Houston, U.S. [Online]. Available: http://www.scholarpedia.org/article/K-nearest_neighbor.

[24] MATLAB 2020a, The MathWorks Inc. ["Silhouette value"]. Natick, Massachusetts, United States; [2020].

[25] Sphinx-Gallery's documents, "Selecting the number of clusters with silhouette analysis on K-Means clustering", Google Summer of Code project, United States, 2007.

[26] Kaufman L., and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons, Inc., 1990.

[27] L. Kaufman, P. Rousseeuw, A. Kassambara, "Finding Groups in Data: An introduction to cluster analysis", John Wiley & Sons, Inc, 1990, p.1-67.

[28] M. Kapildalwani (2015), "Using silhouette analysis for selecting the number of clusters for k-means clustering. (part 2)". [Online]. Available: https://kapilddatascience.wordpress.com/2015/11/10/using-silhouette-analysis-for-selecting-the-number-of-cluster-for-k-means-clustering/

[29] J. D. Rhordes, W. J. Cole, C. R. Upshaw, T. F. Edgar, M. E. Webber, "Clustering analysis of residential electricity demand profiles", in 2014 *Applied Energy journal, Elseiver*.

[30] E. Bobric, G. Cartina, G. Grigoras, “Clustering Techniques in Load Profile Analysis for Distribution Stations”, in *Advances in 2009Electrical and Computer Engineering publication*, Volume 9.

[31] I. Benítez, J.L. Díez, A. Quijano, I. Delgado, “Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance*”, in 2016 Electric Power Systems Research, Elseiver*.

[32] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN", Int. J. Comput. Appl., vol. 105, no. 9, pp. 17-24, 2014.

[33] D. Marutho, S. Hndra, E. Wijaya, “The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News”, *2018 International Seminar on Application for Technology of Information and Communication, IEEE*.

[34] MATLAB 2020a, The MathWorks Inc. [“PCA”]. Natick, Massachusetts, United States; [2020].

[35] F. McLoughlin, A. Duffy, M. Conlon, “A clustering approach to domestic electricity load profile characterization using smart metering data”, in *2015 Applied Energy, Elseiver.*

[36] Iswan, I. Garniwa, “Principal Component Analysis and Cluster Analysis in Profile of Electrical System”, in 2017 *Materials Science and Engineering IOP Conference Series*.

[37] A. Kumar Tanwar, E. Crisostomi, P. Ferraro, M. Raugi, M. Tucci, G. Giunta, “Clustering Analysis of the Electrical Load in European Countries”, University of Pisa, 2017.