POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Machine Learning methods for clustering and day-ahead load forecast of thermal power plants

TESI MAGISTRALE IN ENERGY ENGINEERING – INGEGNERIA ENERGETICA

Scorsolini, Riccardo, 10545944

Advisor:

Prof. Emanuele Giovanni Carlo Ogliari

Co-advisors:

Eng. Alfredo Nespoli

Academic year:

2022-2023

Abstract: The ever-increasing interest in energy efficiency and the onerous amount of data involved makes it necessary to adopt a method that aims to improve the management of a series of buildings connected to a thermal utility. This thesis work aims to provide a method based on unsupervised clustering to group utilities, and then neural networks are used to predict consumption in difficult scenarios of training. The variables that are collected from the smart meters are the basis of the analysis, together with weather parameters. The proposed method has been based on a real working District Heating and therefore, completely real data are used, which are affected by interruptions and missed readings. Therefore, a methodology of data-cleaning, pre-processing and post-processing of data is proposed. The goal of the work is to predict utility consumption through machine learning methods (neural networks) and have been adopted 3 methods for clustering: k-means. hierarchical clustering and DBSCAN. This latter was appropriate for the analyzed case study and hierarchical clustering was found to be the most reliable. Hierarchical clustering was found to be the most reliable and with the most convincing results, according to the indices used to assess the goodness of clustering. In order to predict the required thermal energy consumption, 3 different strategies were adopted, namely: training a neural network for all utilities, neural network one for each cluster, and finally adopting a neural network for each utility. The last methodology is the one that showed better results, but not very far from the second strategy, which could prove to be successful if there is an increase in data and utilities analyzed.

Key-words: Machine Learning, Clustering Methods, Day-Ahead Neural Network Forecast, District Heating

1. Introduction

Access to an ever-increasing range of data inherent in power plant management highlights the importance of using machine learning techniques to aid in the management of the many input data we receive, such as those in the approach to management from the substations of a district heating plant. District heating (DH) represents a valuable way to provide heating and hot water to buildings, through the combined exploitation of renewable sources, industrial excess heat and efficient thermal plants based on fossil fuel [1].

This paper focuses on analysing a methodology suitable for classifying and predicting the energy requirements of a telecontrol network, but extendable to a multitude of applications.

The use of conventional classification and forecasting methods for optimal management of a DH has several critical issues. Indeed, when considering a DH serving many buildings, usually only the data provided by smart meters installed in substations are available. So, using comprehensive physical models is impossible because they exploit data that are unavailable [1].

The strength of this method relies in the fact that only those data that are available from substations are used, which may therefore be affected by missing or errors and that is why careful pre-processing of the data is implemented. Usually, data regarding the variables present are available at a low frequency and for short periods of time, but in our case, data are available on an hourly basis and for more than a year's time. Indeed, thanks to the installation of smart meters, it is possible to have a wide range of monitorable data to make DH control and management more efficient.

Heating load clustering and forecasting can help heating operators meet heat demand in advance and thus formulate wise operation strategies. Currently, there are many studies on DH heat load forecasting [2], such as S. W. Henrik Gadd [3] , in which an introductory analysis is proposed using hourly-based data made available from 141 substations of two networks in Sweden. The same data were then used in study [4] to reveal or failures related to temperature differences and maintaining good quality of the results obtained.

Lu et al. [5] used outdoor temperature, time point, day type and historical heating load of substations as model inputs. Three regression models and the artificial neural network models (ANN) were taken as prediction techniques. The mean absolute percentage error (MAE) was less than 15%.

Another one work where heat load patterns have been analysed for 50 buildings is [6], where the main scope was to estimate heat load capacities for billing purposes. In order to increase energy efficiency in multi-dwelling buildings, heat loads have been monitored and evaluated in [7]. Although the purpose of this work is not primarily to improve energy efficiency in buildings or to estimate heat load capacities but for billing purposes.

Additional investigations, such as that of " *Simple model for prediction of loads in district-heating systems*", have shown how heat demand is predicted in a district heating system, but unlike this work, they rely on a forecast curve for outdoor temperature, this methods instead relies on a historical data set in which other weather measurements also appear [8].

There are various criticisms in using conventional forecast methods for virtual storage applications. Indeed, when DH which serve many buildings are considered, only the data collected by the smart meters installed in the substations (temperature, mass flow rates, and energy consumption) are usually available. This makes the use of full physical models impossible because most data required for complete characterization of the buildings are unavailable [1]. This is one of the reasons why machine learning methods have been used for

data prediction, using Neural Networks (NN). But also, NNs are very well suited for it, for at least two reasons. First, it has been formally demonstrated that NNs are able to approximate numerically any continuous function to the desired accuracy (see [9], [10] and [11] for references).

2. PROPOSED METHODS

The work begins by collecting measured data from substations, after which a data cleaning is carried out. In order to cluster the data, it is necessary to perform an extraction of meaningful features, using the measurements made by meters on substations.

These characteristics will then be correlated, through Pearson's correlation index, with the thermal energy required by the respective substation. For the purpose of clustering, the exogenous characteristics that best correlate with energy demand will be selected.

In addition, available weather data for the site under consideration will be used, because it is widely known that climate has significant influence on building indoor environment and energy consumption [1].

Then the right mix of exogenous and endogenous variables, which thus have the highest Pearson correlation coefficient, will be chosen and forecasting methods, based on neural networks, will be adopted.

Clustering and classification of substation is performed through various techniques, and the results are then compared to demonstrate their goodness.

Clustering refers to a very broad set of techniques for finding subgroups, or clustering clusters, in a data set. When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite like each other, while observations in different groups are quite different from each other [12] .

The clustering task was performed by 3 different methods: k-means clustering, hierarchical clustering and the dbscan method.

Clustering techniques of heat loads based on meteorological variables have been widely used in other studies, such as in the "A clustering-based climatic zoning method for office buildings in China" study. Where thermal load simulations were performed under 274 cities' meteorological conditions across China. Two widely used clustering algorithms: K-Means and Agglomerative Hierarchical Clustering were adopted and compared [13].

Once the clustering part is completed, the thermal demand of the day before and the information extracted from the weather will be used to train the neural networks, with the aim of predicting thermal energy for a given period.

2.1 K-Means clustering:

One of the most common iterative descent clustering algorithms is the k-means algorithm. It's designed for scenarios when all the variables are quantitative [14] .

The algorithm for K-Means Clustering is the following:

• *Assign a number to each observation at random, ranging from 1 to K. These act as the observations' initial cluster allocations.*

• *Iterate the following two points until the cluster assignments stop changing:*

-Calculate the cluster centroid for each of the K clusters. The vector of the p feature means for the observations in the kth cluster is the kth cluster centroid.

- Assign each observation to the cluster with the closest centroid, where closest is defined using Euclidean distance.

The results achieved will be dependent on the initial random cluster assignment of each observation in Step 1 of the Algorithm since the K-means algorithm finds a local rather than a global optimum [12].

K-means method has been implemented for time series clustering. The Matlab's function *kmeans* is used to implement the method. Depending on the type of distance you use there are several possible algorithms to perform clustering: squared euclidean distance, cityblock; cosine, correlation, or hamming [15] .

The Squared Euclidean distance, where each centroid is the mean of the points in that cluster, is calculated as follow:

$$d(x, c) = (x - c)(x - c)' \quad (1)$$

where x is an observation and c is the centroid.

2.2Hierarchical clustering:

The results of applying K-means clustering algorithm depend on the choice for the number of clusters to be searched and a starting configuration assignment. In contrast, hierarchical clustering methods do not require such specifications. Instead, they require the user to specify a measure of dissimilarity between groups of observations, based on the pairwise dissimilarities among the observations in the two groups [14].

The two main categories of methods for hierarchical cluster analysis are divisive (top-down) methods and agglomerative methods (bottom-up), however the agglomerative methods are wider use.

Agglomerative strategies start at the bottom and at each level recursively merge a selected pair of clusters into a single cluster. This produces a grouping at the next higher level with one less cluster. The pair chosen for merging consist of the two groups with the smallest intergroup dissimilarity.

Divisive methods start at the top and at each level recursively split one of the existing clusters at that level into two new clusters. The split is chosen to produce two new groups with the largest between-group dissimilarity. With both paradigms there are N − 1 levels in the hierarchy.

In order to determine how objects in the data set should be grouped into clusters the linkage strategy is used:

• Single linkage, also called nearest neighbor, uses the smallest distance between objects in the two clusters.

• Complete linkage, also called farthest neighbor, uses the largest distance between objects in the two clusters

• Average linkage uses the average distance between all pairs of objects in any two clusters.

•Ward's linkage uses the incremental sum of squares, that is, the increase in the total within-cluster sum of squares because of joining two clusters [16].

The result of hierarchical clustering are precisely hierarchical representations in which each level of the hierarchy generates a cluster that are produced by merging the levels of the previous level. There is only one cluster that collects all the others and at the same time at the basic level each observation has a single cluster.

Each level of the hierarchy represents a particular grouping of the data into disjoint clusters of observations. The entire hierarchy represents an ordered sequence of such groupings. It is up to the user to decide which level (if any) represents a "natural" clustering in the sense that observations within each of its groups are sufficiently more similar to each other than to observations assigned to different groups at that level.

A dendrogram (such as the one shown in Figure 1) is a graphical representation of a highly interpretable full description of hierarchical clustering. The length of the tree's branches along the vertical axes is proportional to the dissimilarity between two clusters. This is one of the primary reasons why hierarchical clustering methods are so popular [8].

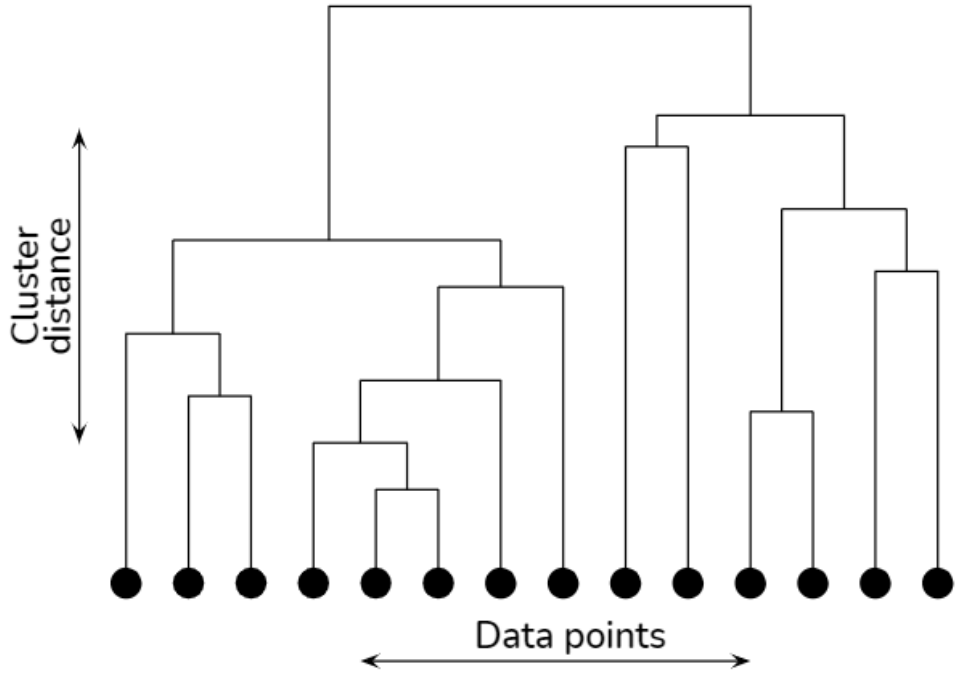


Figure 1- Dendrogram with data points on the x-axis and cluster distance on the y-axis

A dendrogram, as can be seen from the Figure 2, need not branch out at regular intervals from top to bottom as the vertical direction in it represents the distance between clusters in some metric. As you keep going down in a path, you keep breaking the clusters into smaller and smaller units until your granularity level reaches the data sample. In the vice versa situation, when you traverse in up direction, at each level, you are subsuming smaller clusters into larger ones till the point you reach the entire system.

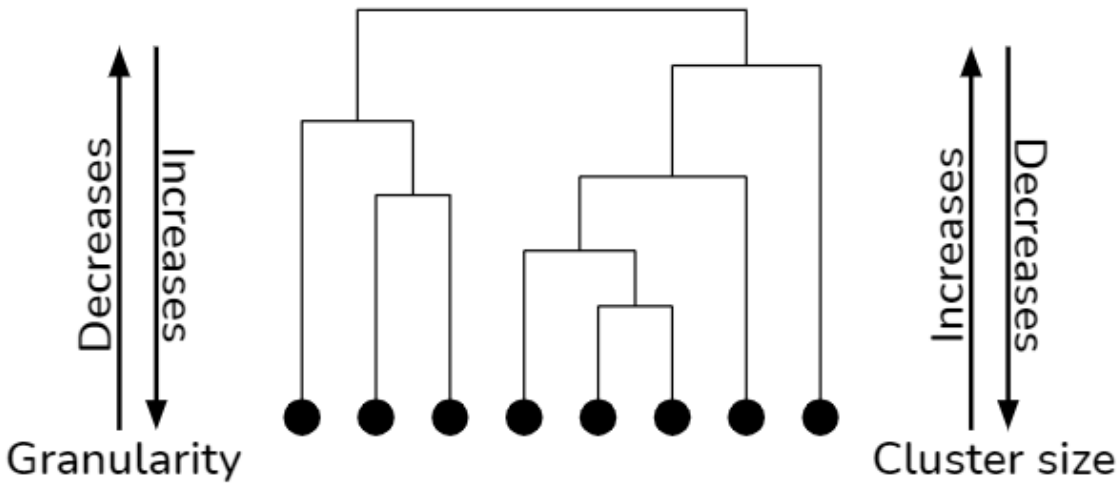


Figure 2- Effect of granularity and cluster size while traversing in the dendrogram

2.3 DB-Scan:

In this section, we present the algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) which is a density-based clustering method and is useful for finding outliers and identifying clusters characterized by an arbitrary shape. designed to discover the clusters and the noise in a spatial database.

The key parameters are:

ε = radius of the considered neighbourhood.

MinPts = minimum number of elements in the neighbourhood to define a cluster.

To find a cluster, DBSCAN starts with an arbitrary point p and retrieves all points density-reachable from p with reference to ε and MinPts. If p is a core point, this procedure yields a cluster with reference to ε and MinPts. If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.

Since we use global values for ε and MinPts, DBSCAN may merge two clusters according to definition *"Let D be a database of points. A cluster C with reference to ε and MinPts is a non-empty subset of D satisfying the following condition of Maximality (∀p,q: if p∈ C and q is a density-reachable from p with reference to ε and MinPts, then q∈ C) and Connectivity (∀ p,q ∈ C: p is density-connected to q with reference to ε and MinPts)"* into one cluster, if two clusters of different density are "close" to each other. Let the distance between two sets of points S1 and S2 be defined as:

$$(S_1 - S_2) = \min\{(p,q)|p \in S_1; q \in S_2\}$$

Then, two sets of points having at least the density of the thinnest cluster will be separated from each other only if the distance between the two sets is larger than ε. Consequently, a recursive call of DBSCAN may be necessary for the detected clusters with a higher value for MinPts. This is, however, no disadvantage because the recursive application of DBSCAN yields an elegant and very efficient basic algorithm. Furthermore, the recursive clustering of the points of a cluster is only necessary under conditions that can be easily detected [17].

2.4 Evaluation indexes:

In order to quantify the goodness of a clustering, indices will be used, which in their complexity provide insight into the validity of the method and the features chosen to perform the grouping.

The clustering goodness-of-fit analysis is evaluated through two indices: the Silhouette index and the Calinski-Harabasz Index. Also, exclusively for hierarchical clustering, the cophenetic correlation coefficient will be used and exclusively for DBSCAN we will assess a priori the parameters of MinPts and ε.

2.4.1 Silhouette Index:

Each cluster is represented by a so-called silhouette, which is based on the comparison of its tightness and separation. This silhouette shows which objects lie well within their cluster, and which ones are merely somewhere in between clusters. The entire clustering is displayed by combining the silhouettes into a single plot, allowing an appreciation of the relative quality of the clusters and an overview of the data configuration. The average silhouette width provides an evaluation of clustering validity and might be used to select an 'appropriate' number of clusters [18].

2.4.2 Calinski-Harabasz Index:

The CH Index (also known as Variance ratio criterion) is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Here cohesion is estimated based on the distances from the data points in a cluster to its cluster centroid and separation is based on the distance of the cluster centroids from the global centroid. Higher value of CH index means the clusters are dense and well separated, although there is no "acceptable" cut-off value. We need to choose that solution which gives a peak or at least an abrupt elbow on the line plot of CH indices. On the other hand, if the line is smooth (horizontal or ascending or descending) then there is no such reason to prefer one solution over others.

2.4.3 Cophenetic correlation coefficient (CPCC):

Hierarchical clustering linked together two data points or objects from the original data set at every level until no objects are there to link. The height of the link illustrates the distance between the two clusters which consists of those objects. This height is known as the Cophenetic distance between two objects. The CPCC values close to 1 are considered as good. CPCC can be used to compare the clustering result of same data set using different distance measures or clustering algorithms. In general, CPCC is a measure of how accurately a dendrogram preserves the pair-wise distances between the time series objects.

Let d_{ij} is the Euclidean distance between the *i*th and *j*th time series objects and $t_{ij}$the dendrogrammatic distance between the two time series objects T_i and T_j. This distance is the height of the node at which these two points are first joined together. Assuming d_{ij}' be the mean of the d_{ij} and t_{ij}' be the average of the t_{ij}, he cophenetic correlation coefficient can be denoted as :

$$Coefficient_{CPCC} = \frac{\sum_{i<j}(d_{ij} - d'_{ij}) \cdot (t_{ij} - t'_{ij})}{\sqrt{\left(\sum_{i<j}(d_{ij} - d'_{ij})^2\right) \cdot \left(\sum_{i<j}(t_{ij} - t'_{ij})^2\right)}}$$

(2)

[19]

2.4.4 Evaluation of MinPits and ε for DB-SCAN:

Before applying the DBSCAN algorithm, a preliminary analysis is performed to determine the best value of MinPts and ε.

To select a value for MinPts, minimum number of elements in the neighbourhood to define a cluster, consider a value greater than or equal to one plus the number of dimensions of the input data. For example, for an n-by-p matrix X, set the value of MinPts greater than or equal to p+1. For the given data set, specify a MinPts value greater than or equal to 4.

To select a value for epsilon, radius of the considered neighbourhood, one strategy is to generate a k-distance graph for the input data X. For each point in X, find the distance to the k-th nearest point, and plot sorted points against this distance. If in the graph there is 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. But before plotting the k-distance graph, it is a good practice to first find the MinPts smallest pairwise distances for observations in X, in ascending order. [20] [17]

2.5 Neural Network Forecast:

Following the clustering phase, the last step of the proposed method is carried out, which is to use neural networks with the aim of predicting the thermal energy consumption for each substation.

Neural Network are mathematical tools inspired by the way the human brain processes information. The basic unit of NNs is the artificial neuron, schematized in the Figure 3. The neuron receives information (which in our case is numerical) through several input nodes, processes it internally, and outputs a response. Processing usually takes place in two steps: first, the input values are combined linearly, then the result is used as the argument of a nonlinear activation function. [9]

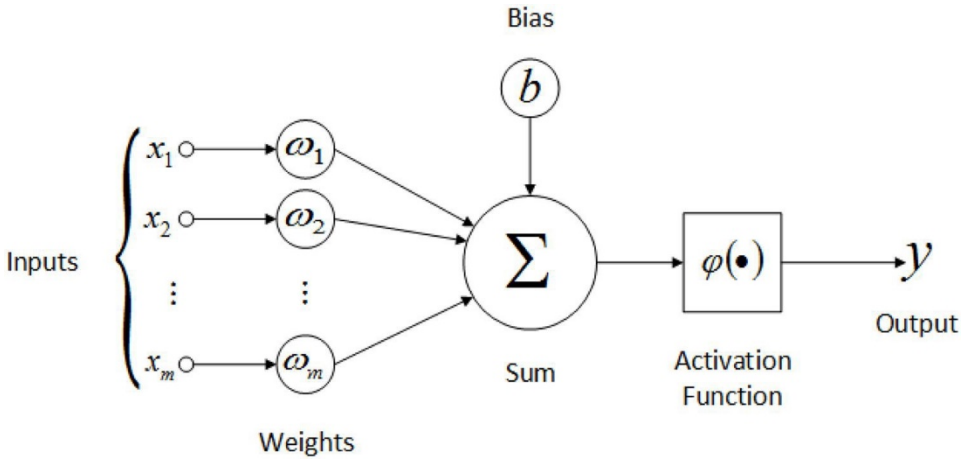


Figure 3 - An artificial neuron

The combination uses the weights ω_i assigned to each connection and a constant bias term b, represented in the figure by the weight of a connection with a fixed input equal to 1. The activation function must be a nondecreasing, differentiable function; common choices are the identity function (y=x) or sigmoid (s-shaped) constrained functions. The architecture of the network is defined based on the arrangement of neurons. Multilayer perceptron (MLP), in which neurons are organized into layers, were used in this work. The neurons in each layer share the same inputs but are not connected to each other. If the architecture is feed-

forward, the outputs of one layer are used as inputs of the next layer. The layers between the input nodes and the output layer are called hidden layers. The Figure belove shows an example of a network with input nodes, hidden layers and one output neuron. [9]

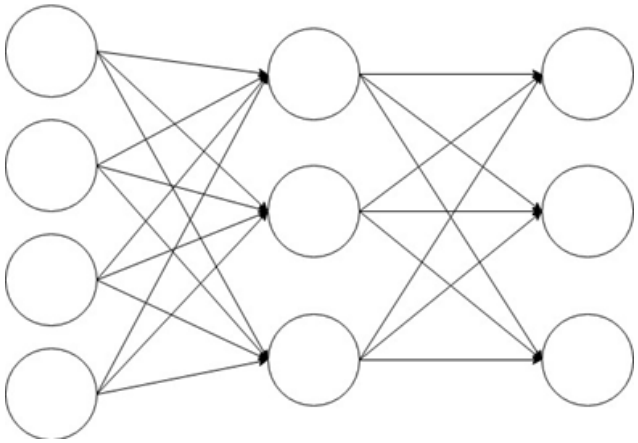


Figure 4 -A two-layer feedforward neural network

The parameters of this network are the matrices of weights (containing the weights connecting the neuron to the input, the weights connecting the output to the neuron) and the bias vector. The estimation of the parameters is called the "training" of the network and is done by the minimization of a loss function (usually a quadratic function of the output error). [9] Basically, training is the process of determining the arc weights which are the key elements of an ANN. The knowledge learned by a network is stored in the arcs and nodes in the form of arc weights and node biases. It is through the linking arcs that an ANN can carry out complex nonlinear mappings from its input nodes to its output nodes [11].

In load forecasting applications, this basic form of multilayer feed-forward architecture shown above is still the most popular [9].

2.5.1 NN evaluation:

After performing the training and validation steps of the neural network, we will go on to calculate the errors committed over the period under consideration, which in our case corresponds to the last week of the dataset. We will then go on to measure the errors made between the predicted and measured thermal energy for the period chosen as a test, which is one week in the following discussion. For each substation we will then have a measurement of:

- Error

$$Err = Forecast\ Energy - Energy\ Test \quad (3)$$

- Mae = Mean Absolute Error

$$Mae = mean(abs(Err)) = \frac{1}{N}\Sigma|err| \quad (4)$$

- Rmse = Root Means Squared Error

$$Rmse = sqrt\left(mean\left(abs(Error)\right)\right) = \sqrt{\frac{1}{N}\Sigma|err|}$$

(5)

- Mbe

$$Mbe = mean\,(Error) = \frac{1}{N}\Sigma Err$$

(6)

3. Case study

The scope of the work is to provide a method in order to cluster and forecast the thermal energy consumption based on an application of a real telecontrol network. The analysis will be conducted based on the data made available by smart meters and those from the appropriate weather station. The district heating consists of 50 substations, enslaved by a power plant, which consist of three boilers and a CHP power plant, based in Chivasso (TO).

3.1 Data collection and Pre-Processing:

The work begins by collecting measured data from substations, after which a data cleaning is carried out. In order to cluster the data, it is necessary to perform an extraction of meaningful features, using the measurements made by meters on substations. Veolia's *Esight* platform was used to collect the data from the meters. Which gives us a set of data that can be measured (M) or calculated (C).

Hourly data from each substation were extracted from the monitoring site. For each substation we have the following measurements available:

- Opening Control Valve [%] - (M).
- Primary delta temperature [°C] - (C).
- Secondary delta temperature [°C] - (C).
- Pressure difference [bar] - (M).
- Heat exchanger efficiency [%] - (C).
- Thermal energy [kWh] - (M).
- Instantaneous flow rate [m3h] - (M).
- Secondary flow rate [m3h] - (C).
- Instantaneous power [kW] - (C).
- Primary supply temperature [°C] - (M).
- Secondary supply temperature [°C] - (M).
- Return temperature primary [°C] - (M).
- Return temperature secondary [°C] - (M).
- Primary water volume [m3] - (M).

The data we have available are on an hourly basis and are available from September 1, 2020, to April 31, 2022.

Looking at the readings provided by the meters we can appreciate that some data are missing, this lack can be attributed to a variety of reasons due to, for example: meter maintenance or lack of signal reception.

For this reason, a cleaning of the data was necessary, opportunely excluding missing data from subsequent analyses.

Since the telecontrol network has intermittent operation, indeed it operates only at certain times of the day and year, especially in the so-called "Thermal Season," which is the period such that heating systems in public places and private homes can be turned on, according to the Italian regulations. At the considered site, the law stipulates that heating systems can be turned on from October 15 to April 15. For this reason, we wanted to create a dataset parallel to the previous one, where the variables are referred only to the period when the demand for thermal energy is greater than zero.

This shifts the analysis only to the actual operating hours of the system, avoiding the times when the demand is zero.

The next stage of data collection is carried out by collecting available data from the meteorological station at the site of interest [21]. Specifically, daily data are available for the entire period under consideration. Those data are:

- Minimum Outside Temperature [°C].
- Maximum Outside Temperature [°C].
- Average Outside Temperature [°C].
- Average Relative Humidity [%].
- Rain [mm].
- Average Wind Speed [m/s].
- Wind Gust [m/s].
- Atmospheric pressure [hPa].

3.2 Feature Extractions

At this stage, the information needed for evaluation was extracted, specifically mean, standard deviation and covalence of all endogenous variables.

After that through the Pearson's correlation coefficient, it was searched which indicator correlated best with the thermal energy demanded by the utilities.

The Pearson's correlation coefficient is measured on a scale with no units and can take a value between -1 and +1. If the sign of the correlation coefficient is positive, it means that there is a positive correlation, otherwise it indicates that there is a negative correlation between the indicators. The closer the value is to +1 or -1, the stronger the correlation between the variable under consideration and the required thermal energy.

The validity of using the Pearson coefficient, as an index for correlation, is supported by various studies based on empirical data. According to researchers that analyzed the relationship between the number of involuntary admissions (detentions) for mental disorders per year under the Mental Health Act 1983 and the number of NHS psychiatric beds each year in England, Pearson's correlation coefficient provides a measure of the strength of the linear association between two variables [22].

The correlation coefficient of two random variables is a measure of their linear dependence. If each variable has *N* scalar observations, then the Pearson correlation coefficient is defined as:

$$\rho\,(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i-\mu_A}{\sigma_A}\right)\left(\frac{B_i-\mu_B}{\sigma_B}\right) \tag{7}$$

where $\mu_A$and σ_A are the mean and standard deviation of A, respectively, and μ_B and σ_B are the mean and standard deviation of B. The correlation coefficient *matrix* of two random variables is the matrix of correlation coefficients for each pairwise variable combination,

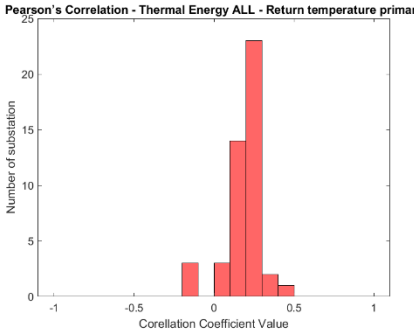
$$R = \begin{pmatrix}\rho(A,A) & \rho(A,B)\\ \rho(B,A) & \rho(B,B)\end{pmatrix} \tag{8}$$

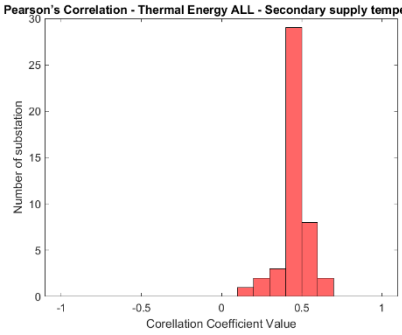
Since A and B are always directly correlated to themselves, the diagonal entries are just 1, that is,

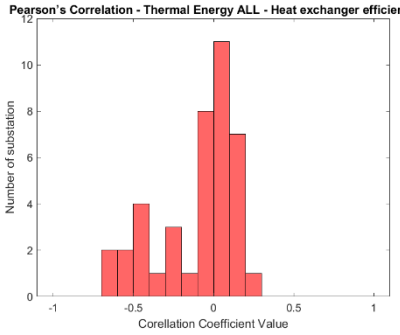
$$R = \begin{pmatrix}1 & \rho(A,B)\\ \rho(B,A) & 1\end{pmatrix} \tag{9}$$

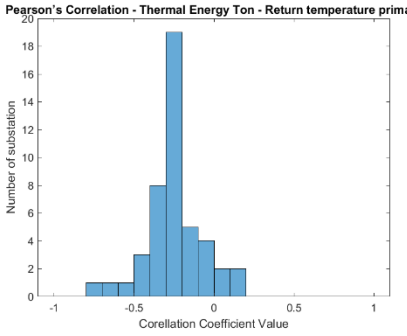
[23]

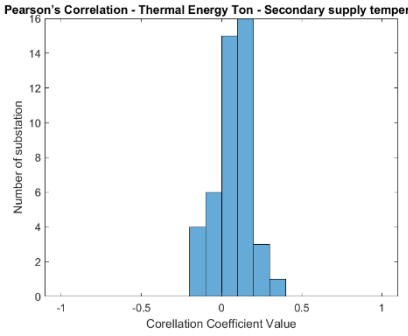
Application of this index between Thermal Energy and the other measured variables produced the following results, it is important to note that given the paucity of samples for *Instantaneous power*, these variables were excluded from the discussion. By calculating the correlation coefficients for the indices, we can see which ones correlate best with thermal energy demand, which will then be the basis of the clustering step. In the following we will use both the hourly thermal energy for the entire period under consideration and using only the values of thermal energy that is actually required, which is when the plant is operating, in our analysis called "Ton". At the beginning was reasonable to calculate Pearson's correlation coefficients only when the power plants was working (i.e. during Ton). However a deeper analysis was lead on Tall providing better results.

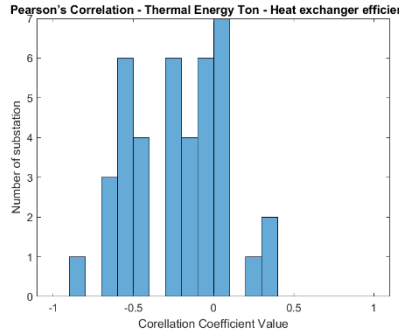

Pearson's Correlation - Thermal Energy ALL - Return temperature primary
Number of substation
Corellation Coefficient Value

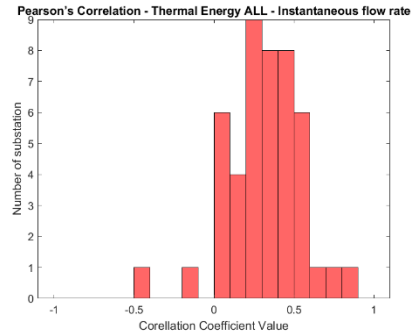

Pearson's Correlation - Thermal Energy ALL - Secondary supply temperature
Number of substation
Corellation Coefficient Value

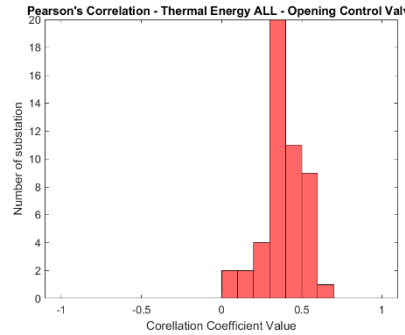

Pearson's Correlation - Thermal Energy ALL - Heat exchanger efficiency
Number of substation
Corellation Coefficient Value

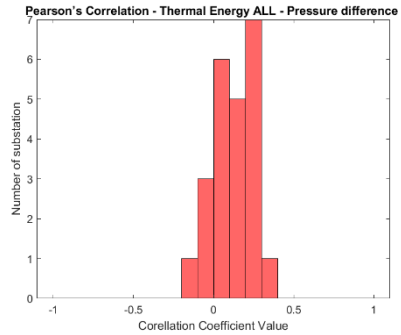

Pearson's Correlation - Thermal Energy Ton - Return temperature primary
Number of substation
Corellation Coefficient Value

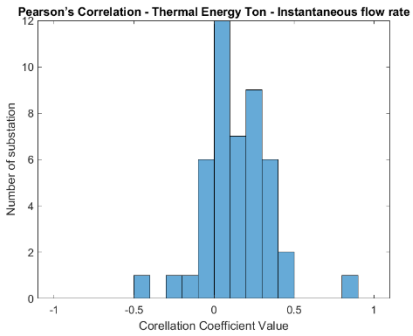

Pearson's Correlation - Thermal Energy Ton - Secondary supply temperature
Number of substation
Corellation Coefficient Value

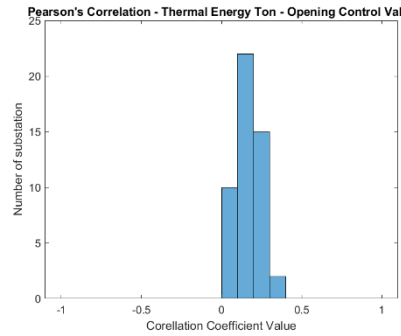

Pearson's Correlation - Thermal Energy Ton - Heat exchanger efficiency
Number of substation
Corellation Coefficient Value

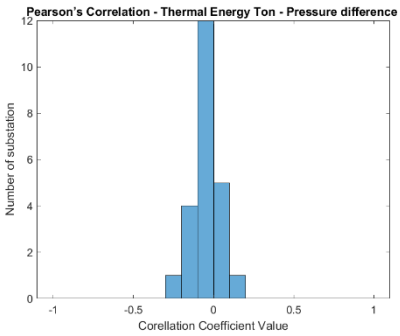

Pearson's Correlation - Thermal Energy ALL - Instantaneous flow rate
Number of substation
Corellation Coefficient Value

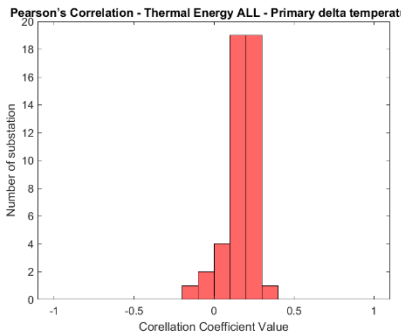

Pearson's Correlation - Thermal Energy ALL - Opening Control Valve
Number of substation
Corellation Coefficient Value

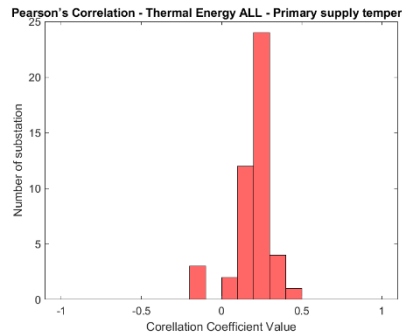

Pearson's Correlation - Thermal Energy ALL - Pressure difference
Number of substation
Corellation Coefficient Value

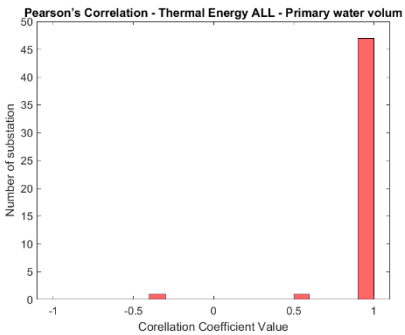

Pearson's Correlation - Thermal Energy Ton - Instantaneous flow rate
Number of substation
Corellation Coefficient Value


Pearson's Correlation - Thermal Energy Ton - Opening Control Valve
Number of substation
Corellation Coefficient Value


Pearson's Correlation - Thermal Energy Ton - Pressure difference
Number of substation
Corellation Coefficient Value


Pearson's Correlation - Thermal Energy ALL - Primary delta temperature
Number of substation
Corellation Coefficient Value


Pearson's Correlation - Thermal Energy ALL - Primary supply temperature
Number of substation
Corellation Coefficient Value


Pearson's Correlation - Thermal Energy ALL - Primary water volum
Number of substation
Corellation Coefficient Value

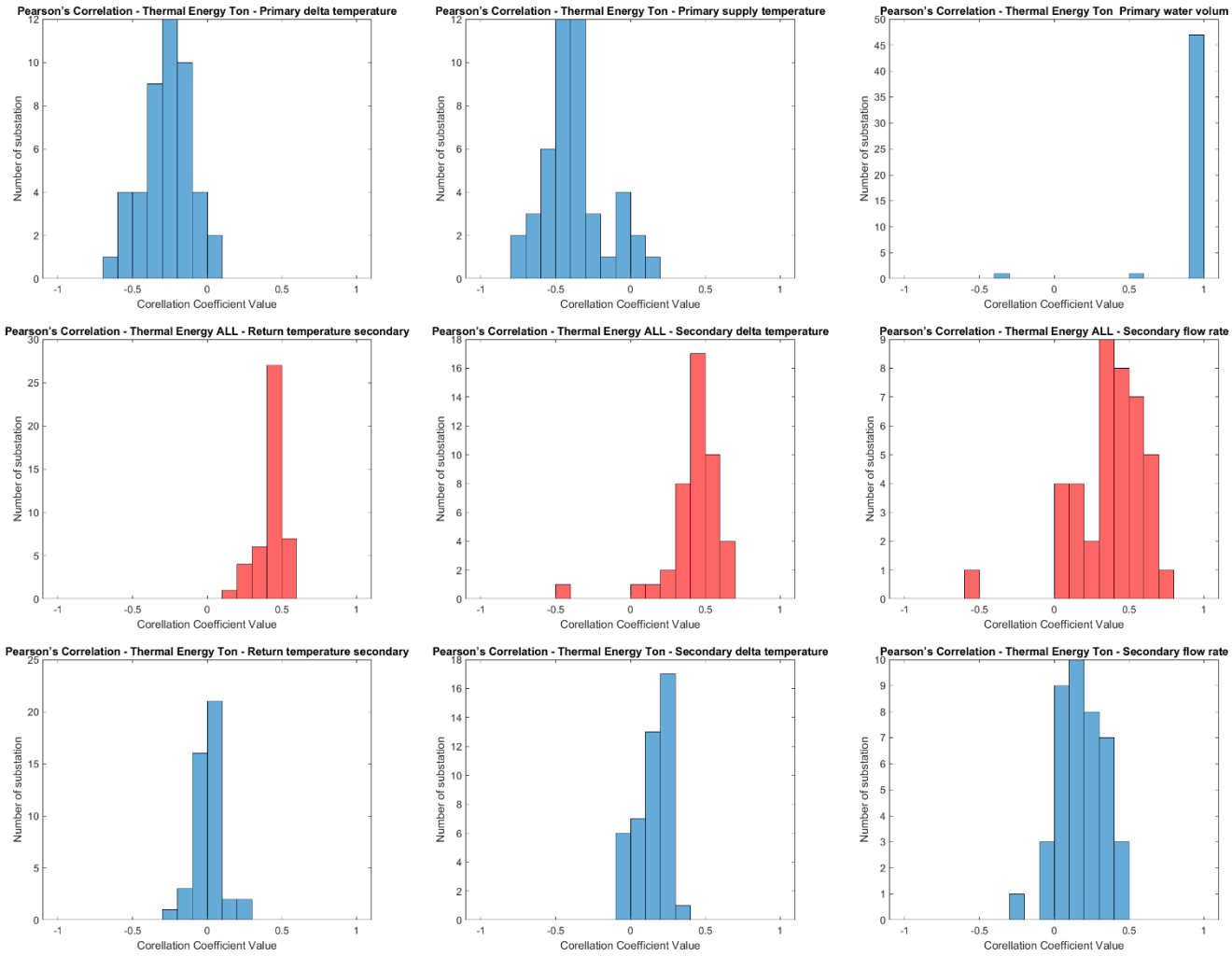


Figure 5 - Correlation between Thermal Energy and Measured Variables

Therefore carrying out the correlation coefficient analysis between Thermal Energy and the other variables for all plants, the average over 41 plants was taken, the values of which are in the following Table 1.

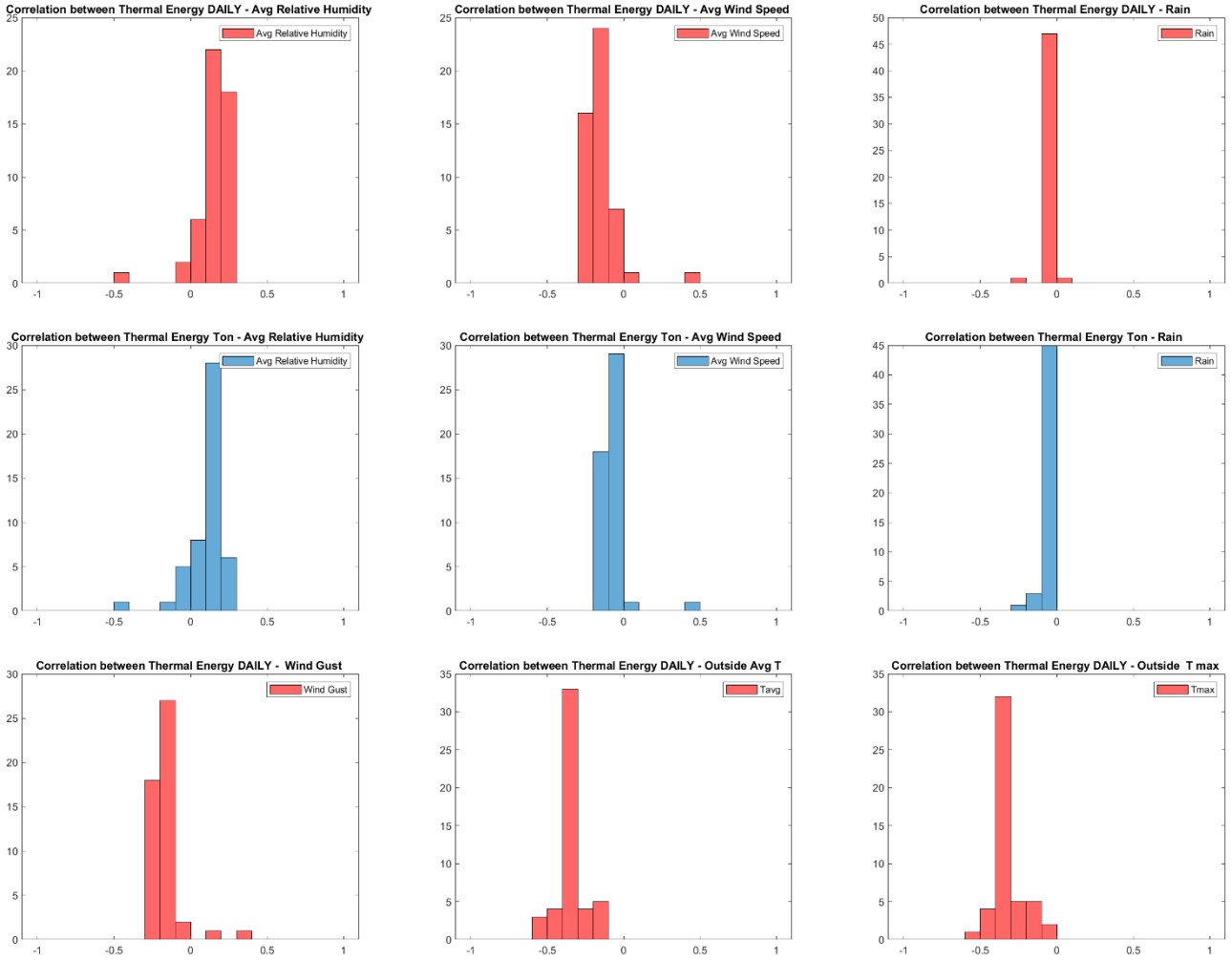
Table 1- Pearson's Correlation Coefficient between Thermal Energy required and Variables

	Mean Pearson's Correlation Coefficient Ton	Mean Pearson's Correlation Coefficient ALL
Opening Control Valve	0.172	0.386
Primary delta temperature	-0.263	0.167
Secondary delta temperature	0.147	0.422
Pressure difference	-0.057	0.123
Heat exchanger efficiency	-0.229	-0.098
Instantaneous flow rate	0.140	0.322
Secondary flow rate	0.178	0.379
Primary supply temperature	-0.380	0.196
Secondary supply temperature	0.074	0.456
Return temperature primary	-0.254	0.194
Return temperature secondary	0.000	0.425
Primary water volume	0.952	0.953

It is evident that in our case study the use of variables for only the actual period of thermal energy demand it does not result in a significant increase in correlation coefficients and indeed in some cases even decreased

them. For this reason, in the next step of clustering, it was deemed appropriate to use the variables calculated over the entire time period (Tall).

During the investigation it was also possible to identify which variable exogenous to the system had an appreciable correlation, so that it could be used as a distinguishing parameter in the various analysis. In particular, the analysis on exogenous variables looked at weather parameters available at the Chivasso weather station, with daily data, using the same Pearson coefficient. Histograms are presented below which indicates how each substation's Thermal Energy correlates with the corresponding weather variable, calculated daily and then averaged. The analysis was conducted both by taking the entire day as the parameter for Thermal Energy, in red, but also by considering the exact hours of operation, in blue.



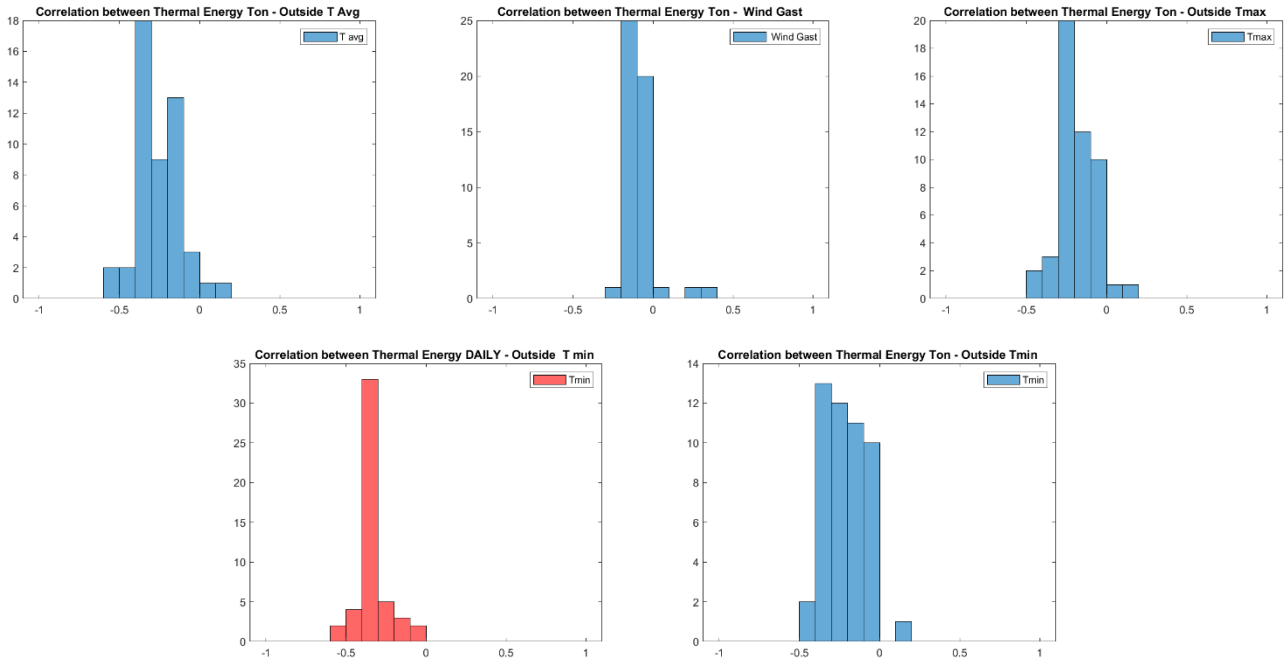


Figure 6 -Correlations Between Thermal Energy and weather indexes

Looking at the histograms of the weather correlation coefficient, we can see that the use of the Thermal Energy Required only in the on times (those colored blue) increases little the correlation coefficient between the weather variables and Thermal Energy. In particular, the rise, in absolute value, is marked for Average, Maximum and Minimum outside Temperatures; a slight increase is also appreciable for the variables of Rain, Average Relative Humidity and Wind Gusts. In contrast, the correlation worsens slightly with the variable corresponding to Mean Wind.

The next analysis considers using other weather stations at greater distances from the generating station and see how the Pearson's correlation coefficient varies. We can thus appreciate how the method used is affected by using weather data that are farther away from the site of interest. Referring to the following weather stations: Caselle (TO) at 18 km, Bric della Croce (TO) at 21 km, Altessano (TO) at 22 km and Collegno (TO) at 30 km(Figure 7). This analysis was carried out to observe how affected they are if weather data were available that were not precisely referenced to the site under consideration.

Figure 7 - Overview of different weather stations

By repeating the same procedures as before we obtain the histograms that are represented the below.

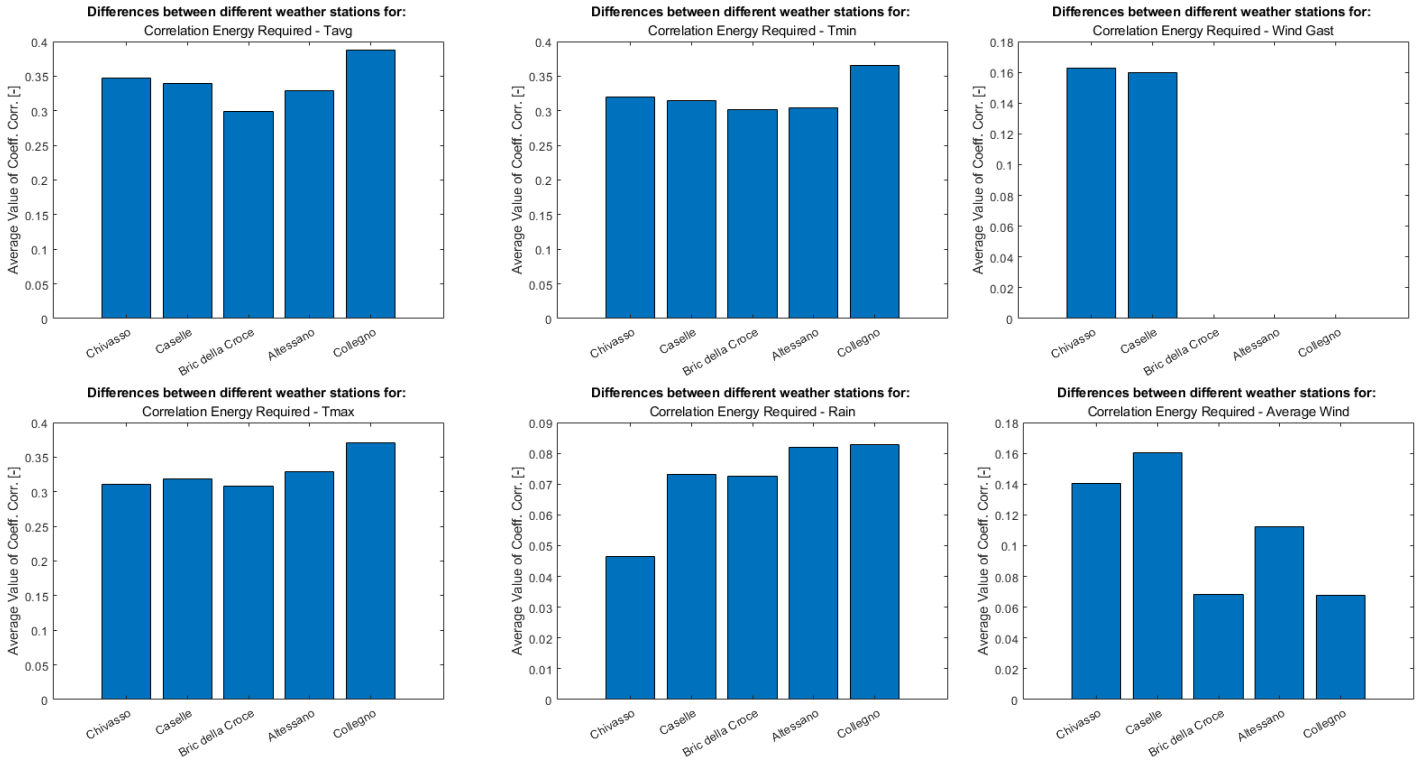


Figure 8-Differences in Correlation Coefficient among different weather stations

Looking at the results obtained Figure 8, we can see that even if it is generally preferable to use as the weather station the one referring to the site under consideration, but if no weather data are available, neighbouring weather stations can also be used with good confidence.

3.3 Clustering

The clustering activity is carried out by seeking which parameters correlate best with thermal energy. Looking at the correlation coefficient values, we can see that the variables that have higher correlation index with thermal energy demand are:

- Opening Control Valve – 0,386
- Secondary delta temperature - 0,422
- Secondary flow rate – 0,379
- Return temperature secondary – 0,425
- Secondary supply temperature – 0,456
- Primary water volume - 0,953

Of the following parameters, we chose not to use the values of *Secondary delta temperature* and *Primary water volume*, because they are derived from calculations or affected by other regulation logic for the entire district heating network. Thus, for our purposes not appropriate. Therefore, all clustering algorithms were tested using several parameter combinations based on the previous analysis. It was chosen to operate with three measurements at a time for each substation. After numerous tests and subsequent analysis, based on the clustering goodness-of-fit indices, the combination that guaranteed the best cluster partitioning in our case study was the one that used:

- *Average Thermal Energy*.
- *Average Secondary supply temperature.*
- *Standard deviation of Instantaneous flow rate.*

3.4 Neural Network Forecast

After clustering the data, the method considered proposes the following strategies so that we can investigate which one best fits the case study. In order to make the best use of the neural networks, new data pre-processing was carried out to make the best use of the available hourly data. All plants that had data available only from October 2021 were removed, and the part of the dataset that included periods outside the "Thermal Season" was removed. This strategy was adopted in order to best train the neural networks only in the actual period of use, not affecting their analysis, since in the period outside the DH is actually off. In addition, all missing hourly data gaps were interpolated based on data from the previous and next 2 hours, to provide to the NN with as homogeneous a database as possible. After making this data adjustment, the database on which to perform our analysis saw the number of substations reduced to 33 and the hourly data was 6137 samples.

Consulted weather data on an hourly basis from the reference weather station was then used. From that meteorological station, however, data from 2021 were not available on an hourly basis, so available values from the typical meteorological year [24] were taken. The weather data provided as input to the network are thus:

- Outside mean Temperature [°C].
- Wind speed [m/s].
- Pressure [hPa].
- Relative Humidity [%].

To train the network to predict the next day's thermal consumption, available weather values and the thermal energy demanded at the same time on the previous day were provided as inputs. Regarding the network training, three different strategies were adopted, which are:

1. Provide as training input the entire dataset of the 33 substations, except the last week, which is used as a target and on which the goodness of the neural network will be tested.
2. Train one neural network per cluster for the entire time frame of the dataset, except for the last week, which is used as the target and on which the goodness of the neural network will be tested.
3. Train a neural network for each individual substation for the entire time span of the dataset, except for the last week, which is used as the target and on which the goodness of the neural network will be tested.

Regardless of the strategy used and the inputs, the neural network that is used in this work is characterized as follows:

- The dataset is divided into two parts: input and target. The input consists of the required heat energy available for the entire dataset except the last week, which is used as the target. The input is then further divided randomly, where 70% of the data is used as training and 30% is used as validation.
- Training was done by conducting 10 trials in which data were always randomly divided between validation and training.
- 20 hidden layers were chosen for the neural network.
- The training method chosen is the Levenberg-Marquardt.

3.4.1 Neural Network Post-Processing

After processing the data through the neural networks, post-processing of the data was performed. Particularly during the night-time hours, the predicted thermal energy was very low, while the actual thermal energy was zero. For this reason, a constraint was imposed on the predicted energy, which is set to zero when the measured energy goes to zero. This choice is justified by the fact that indeed the district heating network has a very rigid operating range, and this forcing does not vary much what the neural networks produce. It turns out to be only a formal correction based on DH operation.

3.5 Proposed Methodology:

In conclusion, the methodology expressed so far is intended to provide a method that is as flexible and adaptable to the case study as possible, so that it can be used and achieve useful and effective results. What this work aims to do is to provide a way forward if one intends to classify consumers and forecast thermal energy consumption within a district heating network. Where the ultimate goal of the methodology is to provide information that is useful and as reliable as possible so that resources can be managed efficiently and with rational use. The proposed methodology followed the following steps.

Figure 9 - Methodology Flowchart

4. Results:

The analysis was performed using Matlab algorithms of related clustering techniques and for the use of neural networks. In particular, the Matlab's Statistics and Machine Learning Toolbox package was used. So, following all the steps in the methodology outlined and applying the strategies listed above yields the following results.

4.1 Clustering Results:

4.1.1 Remove Outlier:

Before implementing the clustering techniques, an investigation of the presence of outliers that would affect the final result and the correct representation of them was carried out. Specifically, once the right combination of information to be used for classification was identified, a boxplot was developed in order to observe the presence or absence of any outliers. As we can see from Figure 11, outliers are present and that affect the correct grouping. In particular, the plants that are out of range are the same for all the variables examined. As proof of this, if we look at the dendrogram (Figure 10) carried out without the removal of outliers, we can see that they negatively affect the distribution and thus increase the number of clusters incorrectly. In fact, each outlier represents a cluster with only one member. For these reasons, they were excluded from the next stage of clustering and treated precisely as outliers.

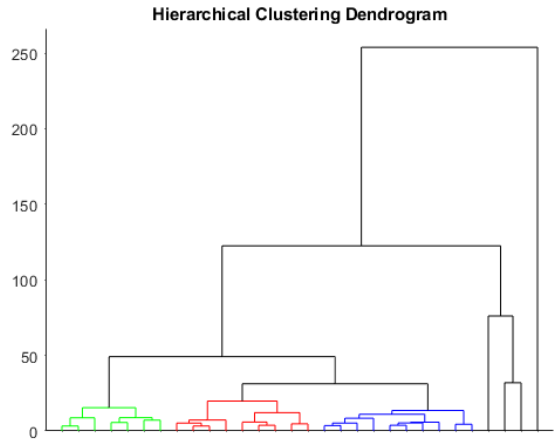


Figure 10 - Hierarchical dendrogram before outliers removal

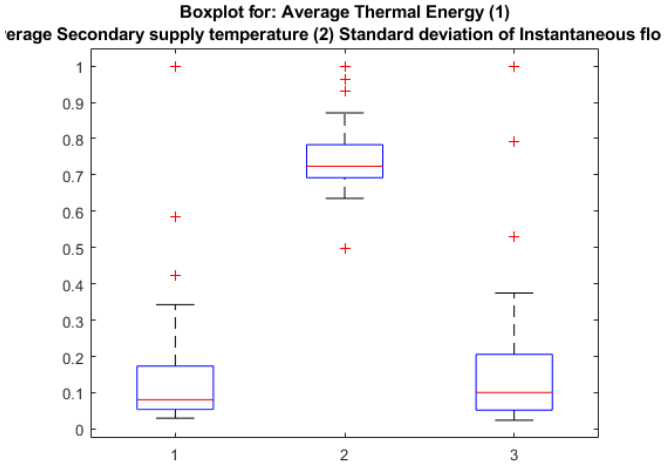


Figure 11 - Boxplot for clustering variables, with outliers plants

4.1.2 K-means Clustering Results:

The implementation of k-means clustering start using the *Squared Euclidean distance* and perform the algoritm with several numbers of clusters: from 2 to 10. Application of kmeans is performed by using *Average Thermal Energy*, *Average Secondary supply temperature* and *Standard deviation of Instantaneous flow rate* as clusering variables. This set of variables was used because after several trials and subsequent analyses, evaluating both

Pearson's correlation coefficients and clustering goodness-of-fit indices such as Slihouette's graphical index and CH index, these were found to provide the most uniform distribution.

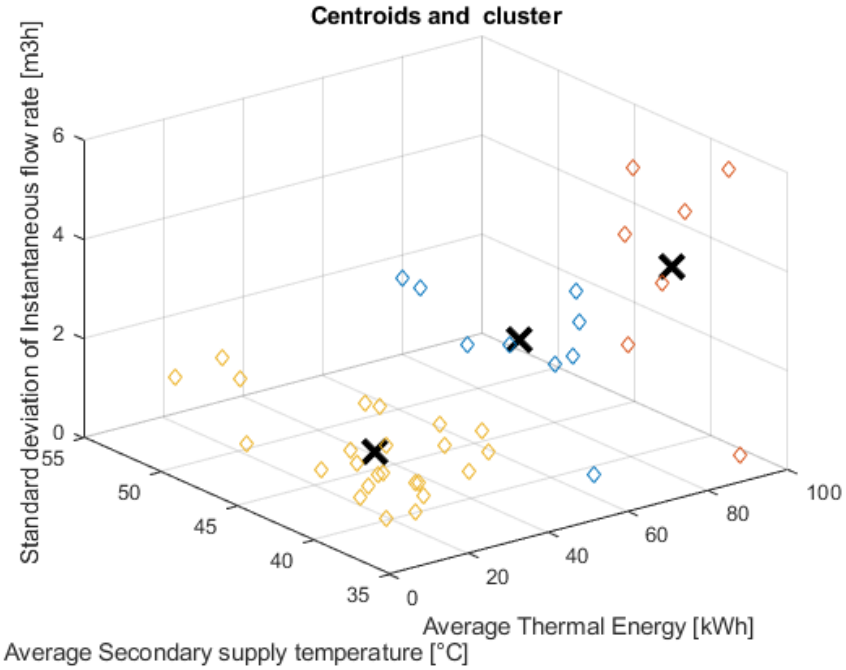


Figure 10 - k-means Clustering Spatial Representation

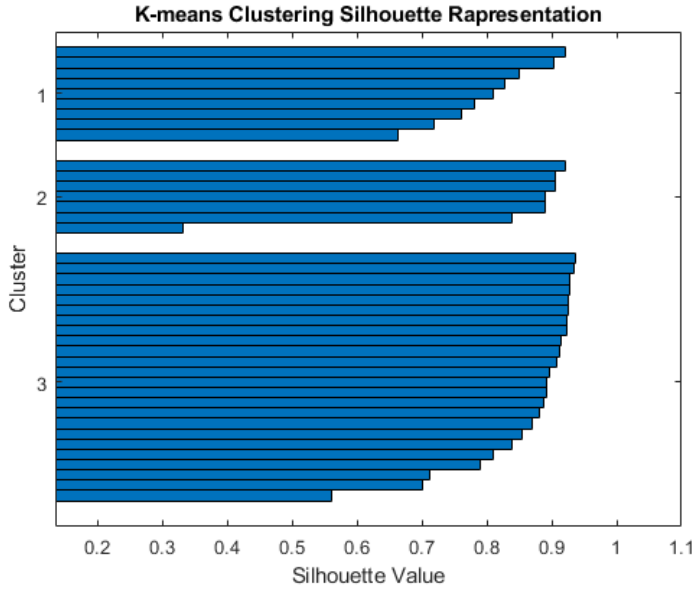


Figure 11 - k-means Silhouette Representation

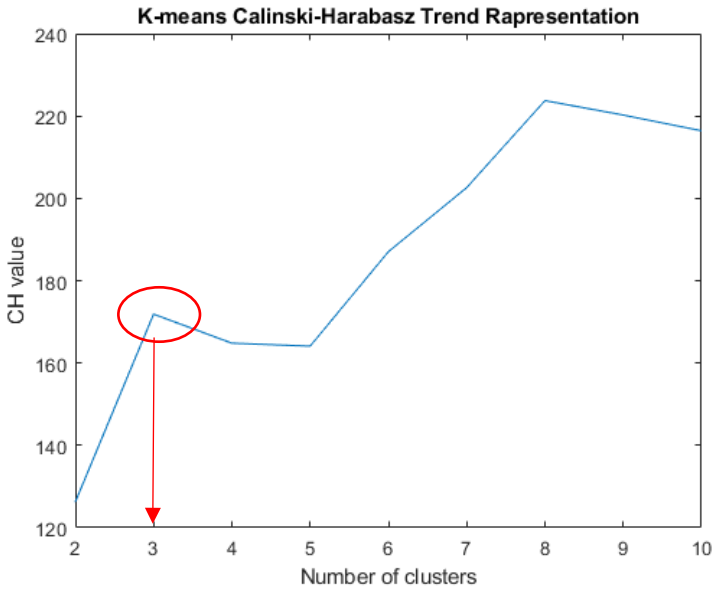


Figure 12 - K-means Calinski-Harabasz Trend Rapresentation

Looking at Figure 12 we can see that the clustering is pretty omogeneous, indeed even the Silhouette index (Figure 13) show that the distinction between the clusters is quite clear. In addition, Figure 14 shows that the number of groups is consistent with the CH trend, that suggests us the optimal number of clusters depending on where we have a negative inflection point, which in our case occurs precisely with the 3 clusters. Although one of the three clusters turns out to be more populous than the others. This limitation is due to the application limitations of the k-means algorithm, in that it cannot efficiently separate some clusters that are not spherical clusters with similar numbers of elements, in besides the fact that the number of clusters must be chosen a priori. For this reason, we need to use a clustering technique that is more appropriate to deepening our analysis.

4.1.3 Hierarchical Clustering Results:

To perform hierarchical clustering, the criterion chosen for defining clusters is *'distance'*, where the cluster groups all substations at or below the level of a node, under the condition that the node height is less than Cut-off. The linkage function takes distance information and links pairs of neighbouring objects into binary clusters, using the Euclidean square distance. Considering *Average Thermal Energy*, *Average Secondary supply temperature* and *Standard deviation of Instantaneous flow rate* as clustering variables, we obtain the following results, where the clustering goodness-of-fit analysis is evaluated by looking at the cophenetic index and looking at the dendrogram with a value of cut-off of: *30*.

With this configuration we obtain 0. 8334 as a cophenetic index, which is a good value since it is close to 1.

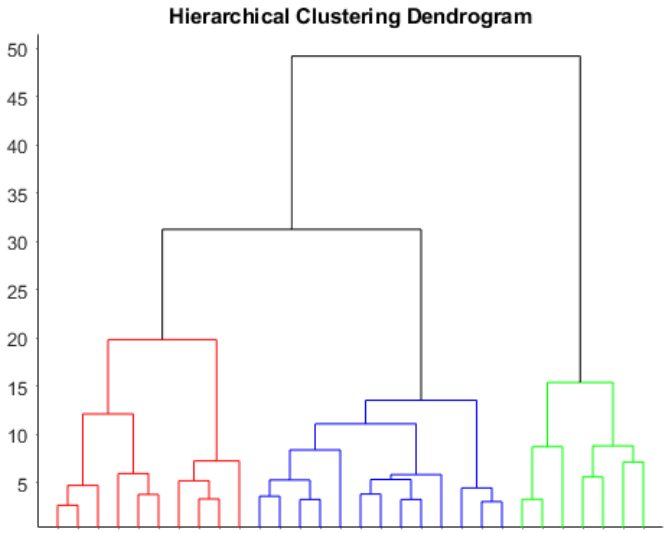


Figure 13 - Hierarchical Dendrogram

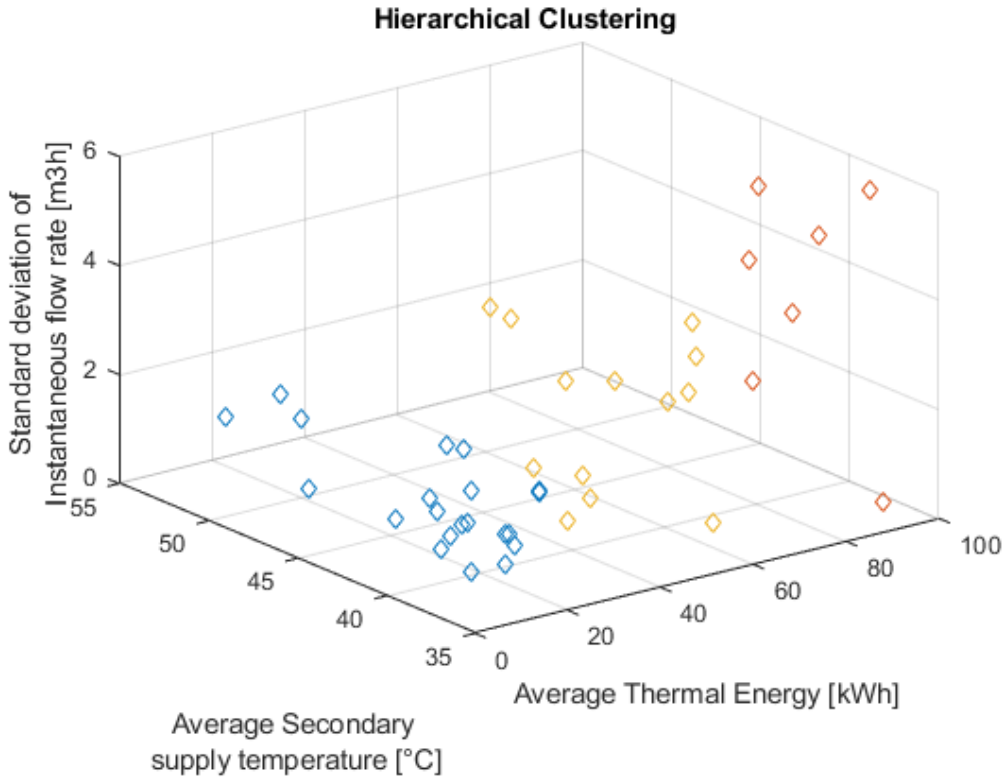


Figure 14 - Hierarchical Clustering Spatial Representation

As we can appreciate from both the dendrogram (Figure 15) and the spatial arrangement (Figure 16), the arrangement of the three clusters is clearly evident. We can also see that one cluster is much more crowded than the other two. To compare the goodness of clustering in addition to the high cophenetic coefficient, we can look at the Silhouette index. In Figure 17 we can appreciate how clusters 1 and 2 are homogeneously populated, while number 3 is almost completely homogeneous except for the elements with negative Silhouette value. For the other clusters we can appreciate the goodness of clusters since all elements possess a good silhouette value, i.e., as close as possible to 1 and above 0.5. Excepts for the elements that hardly fit into the cluster 3, overall, however, all elements are correctly arranged.

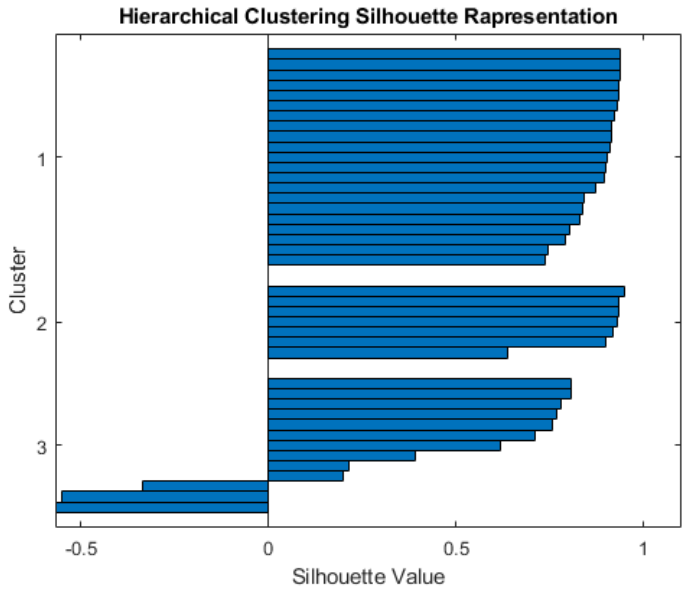


Figure 15 - Hierarchical Clustering Silhouette Representation

4.1.4 DB-SCAN Clustering Results:

Before applying the DBSCAN algorithm, a preliminary analysis is performed to determine the best value of MINPITS and EPSILON. In order to select the number of minpits we look at the size of the matrix of measurements that is used as the variable set for clustering. Since there are N substations in our case study and 3 variables are chosen, the size of the matrix is 3xN. So, a good value for minpits is 4. [20]

To find the value of epsilon instead, we start from the k-graph at its knee point. On subsequent analyses, however, this value was reduced to unpack large clusters, as visible in our case. Hence the optimal value chosen for the case study under consideration was 9.

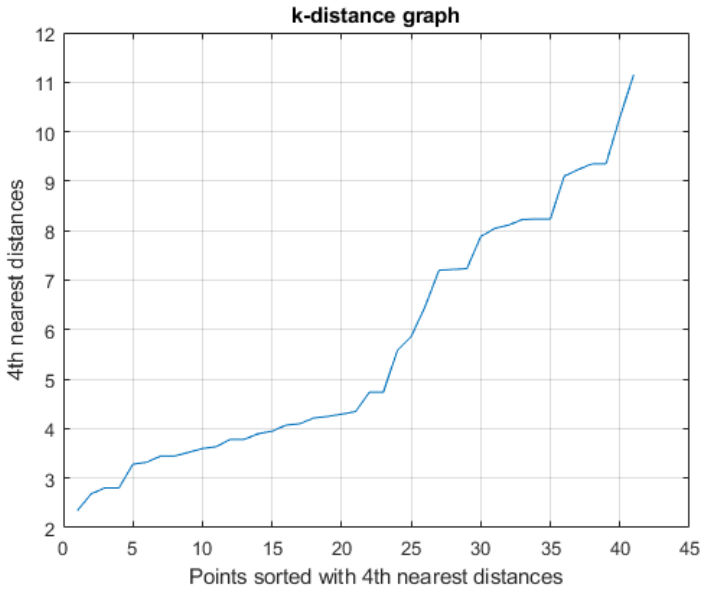


Figure 16 - DBSCAN k-distance graph

Although Figure 18 suggests starting with a minpits value of about 45, after several trials it was decided to decrease this value to that used to break up the larger clusters. In fact, using the value chosen we obtained what is shown in Figure 19.

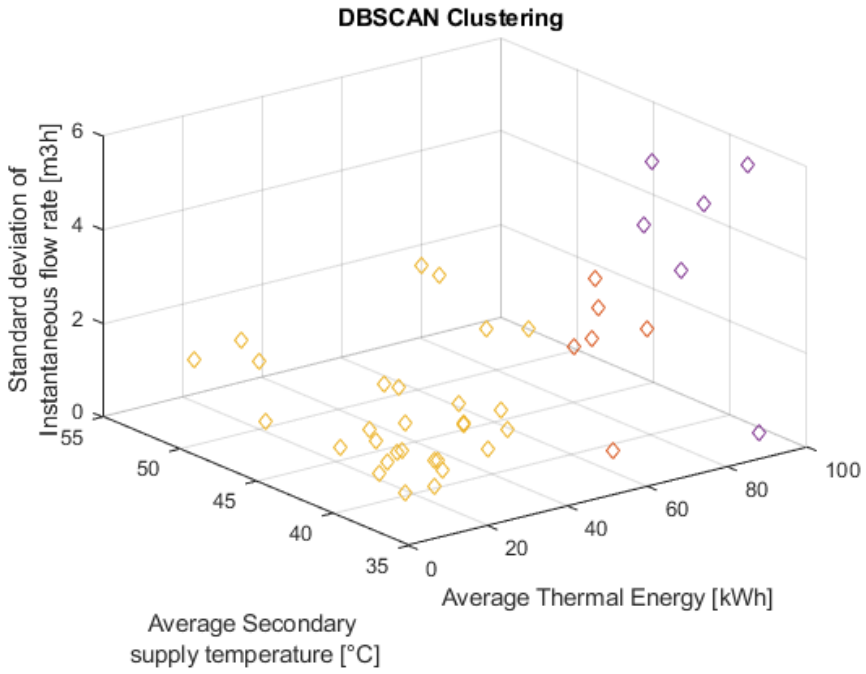


Figure 17 - DBSCAN Spacial Representation

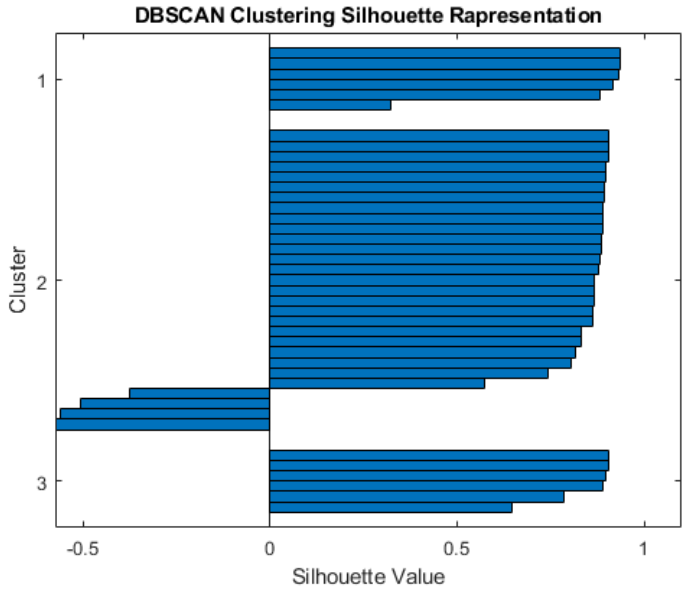


Figure 18 - DBSCAN Silhouette Index Represemtation

What emerges from looking at the DBSCAN results is that the clusters stand at 3 it is noticeable that cluster number 2 contains some elements that are not entirely homogeneous with other members of the same group. Indeed, as can be appreciated in the Silhouette graph, indices are affected by negative values. This representation confirms what has already been obtained in hierarchical clustering. That is, the presence of 3 homogeneous clustering is confirmed, with the presence of some member that have dissimilarity such as those in cluster 2. However, the grouping performed appears to be less efficient than that achieved by the hierarchical method; in fact, DBSCAN merges some of the members into a large group.

4.1.5 Clustering Results:

In conclusion we can state that the number of clusters is 3, one of which is more populated and the other two of those populated almost equally. The sizes that can slightly vary depending on the method indeed both DBSCAN and hierarchical clustering report a similar distribution of clusters except for a few cases on members that are located at the boundaries of the clusters and are attributed to one or the other class according to the grouping criterion.

The result of the unsupervised clustering phase according to what was produced is a breakdown into 3 groups arranged as follows (Table 2), with the presence of 4 outliers.

Table 2 - Cluster Ripartition

	Cluster 1	Cluster 2	Cluster 3	Outliers
Members	21	7	13	4

4.2 Forecasting Results:

Once the clustering phase is completed, further cleaning of the dataset is performed as already described in section 3.4, until a series of 6137 hours is obtained for 33 substations. This hourly series was obtained not only by eliminating all the holes and missing readings, but also all the values equal to 0 that are outside the thermal season, which is the one in which the thermal energy demand is zero by law. The obtained dataset is then given as input to the neural networks according to three different strategies, the purpose is to investigate which method is the most appropriate for our case study based on the errors made in predicting the energy demand for the next 7 days. For prediction, the neural networks are given hourly data from the site weather and the thermal energy demanded the day before the same time that is to be predicted. The results that are obtained under the different strategies are listed below.

4.2.1 Results for one Neural Network for All the Substation – Strategy 1:

Training a neural network for all substation and then going to measure the errors made by the prediction compared to the test, we obtain the results in the Table 3.

Below (Figure 21) we can observe the prediction profiles of each substation, comparing the forecasted to the measured energy requirement in the selected period.

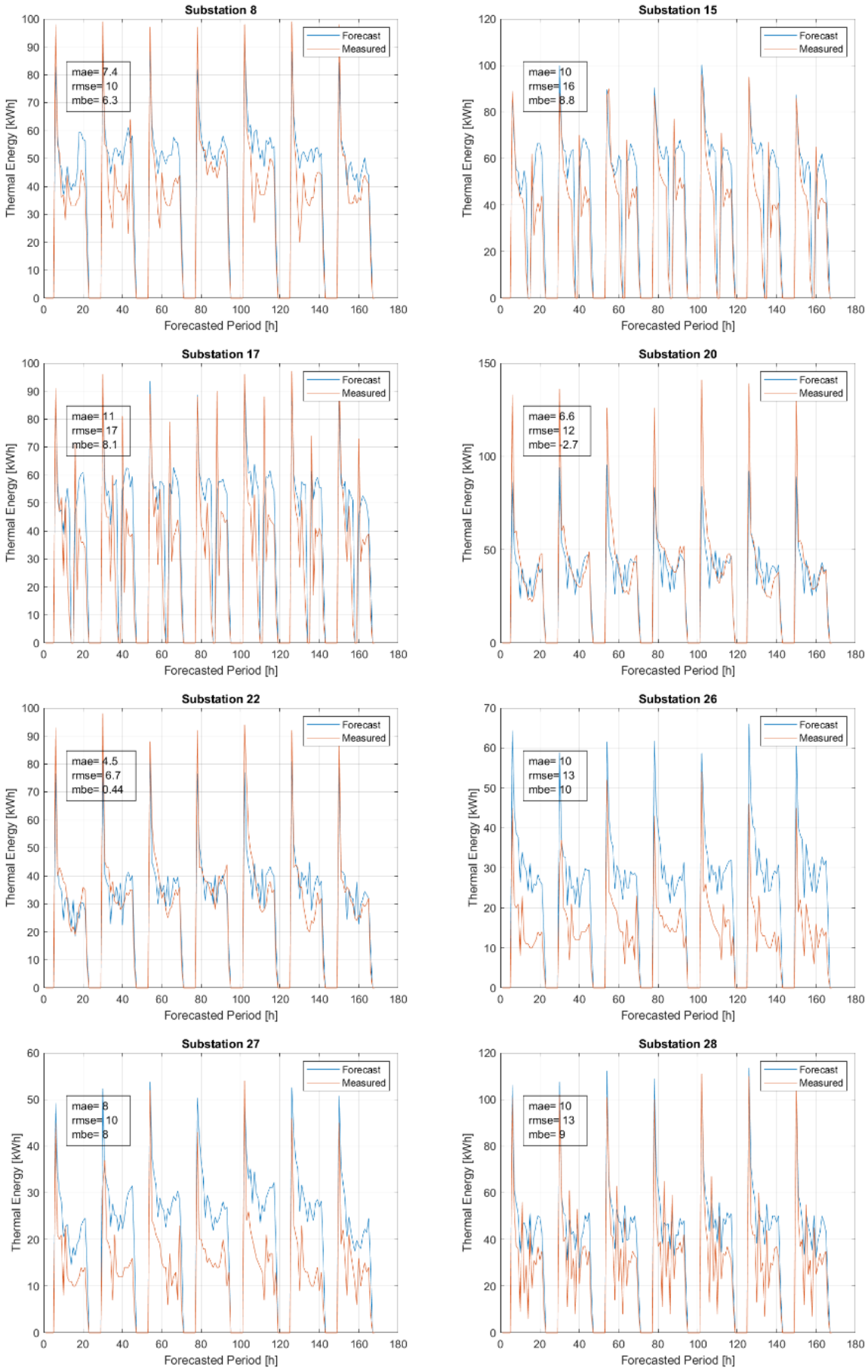


Figure 19 - Forecasted Energy trends compared to measured energy trends for each substation using strategy 1

4.2.2 Results for one Neural Network for Each Cluster– Strategy 2:

Training a neural network for each cluster and then going to measure the errors made by the prediction compared to the test, we obtain the results. Below we can observe observe the prediction profiles for each cluster, comparing the forecasted to the measured energy requirement in the selected period (Figure 22,23,24,25).

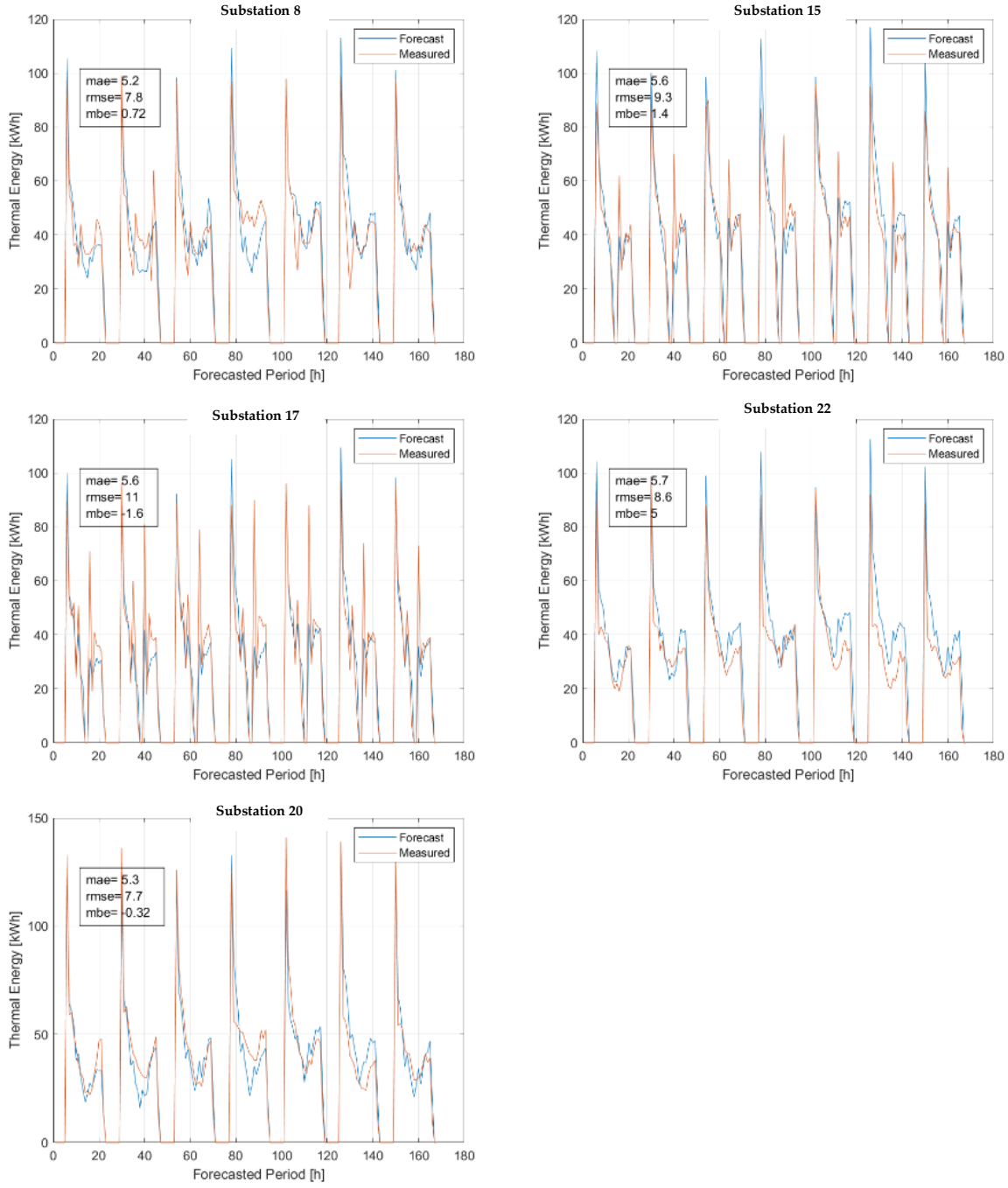


Figure 20 - Forecasted Energy trends compared to measured energy trends for each substation in Cluster A using strategy 2

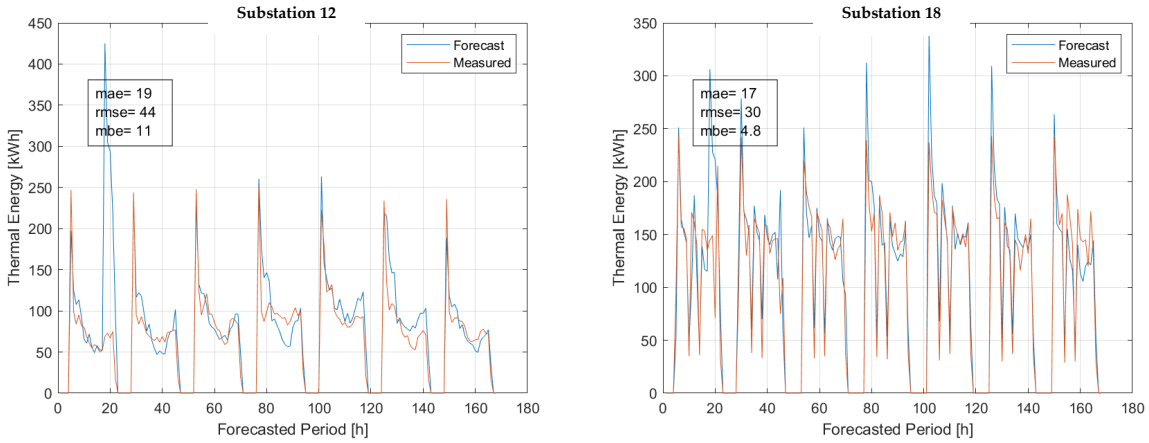


Figure 21 - Forecasted Energy trends compared to measured energy trends for each substation in Cluster B using strategy 2

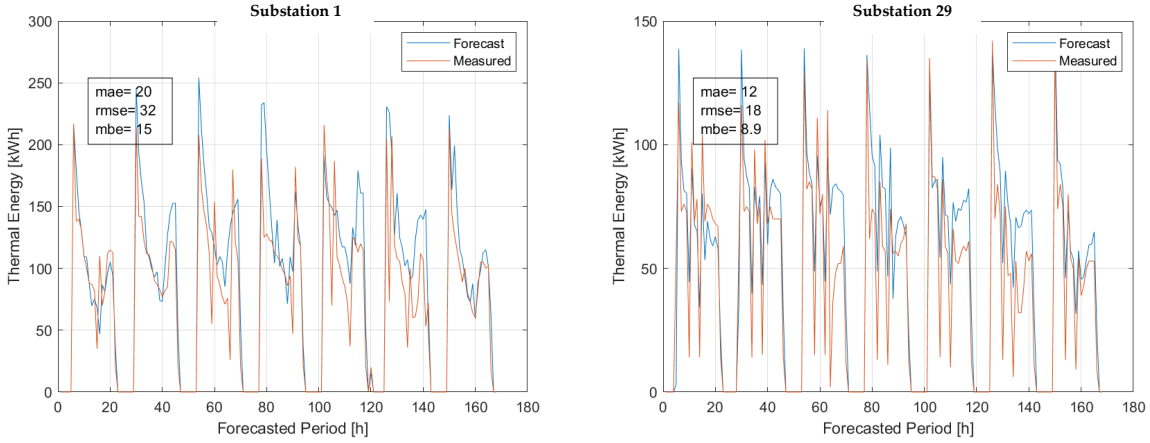


Figure 22- Forecasted Energy trends compared to measured energy trends for each substation in cluster C using strategy 2

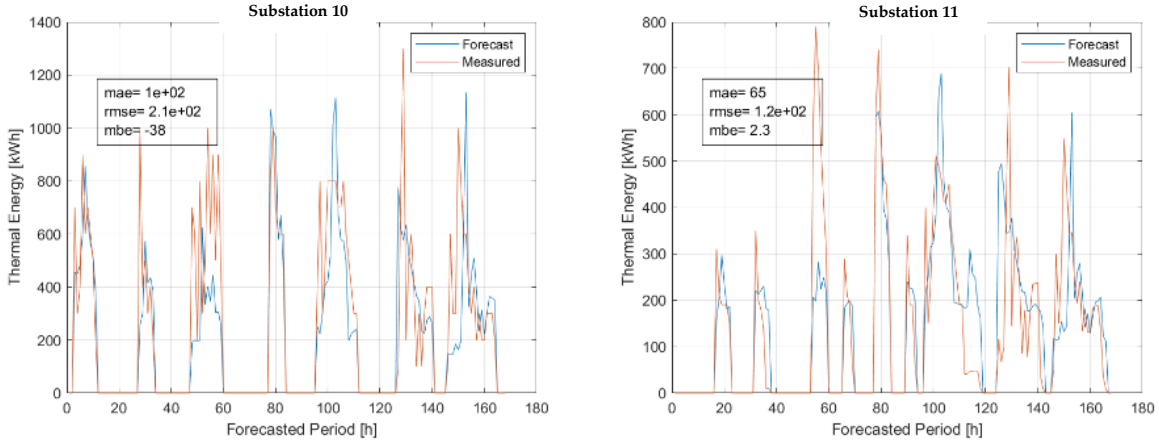


Figure 23 - Forecasted Energy trends compared to measured energy trends for each substation for the Outlier using strategy 2

4.2.3 Results for one Neural Network for Each Substation– Strategy 3:

Training a neural network for each substation and then going to measure the errors made by the prediction compared to the test, we obtain the results showed in Figure 26.

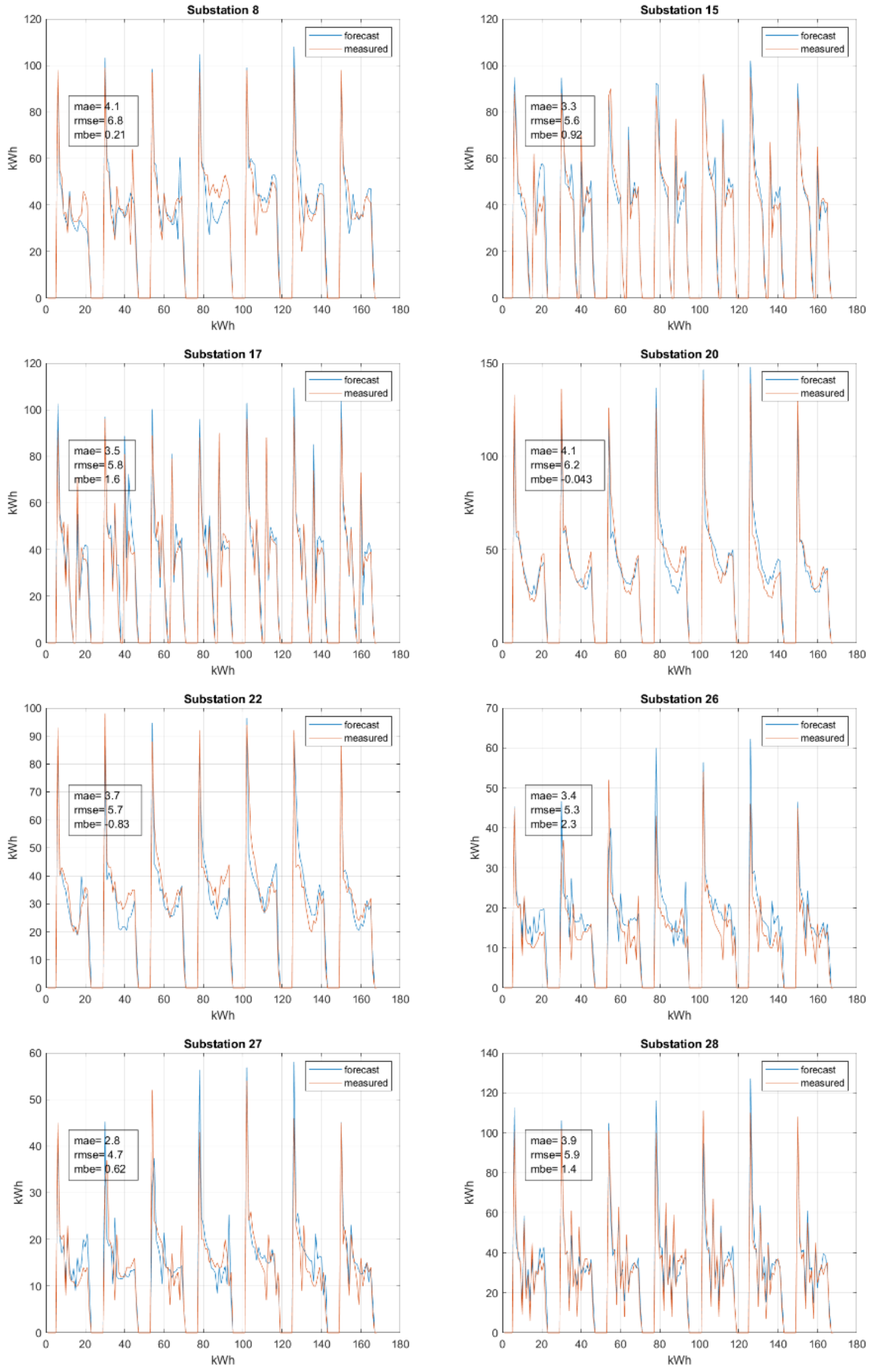


Figure 24 - Forecasted Energy trends compared to measured energy trends for each substation using strategy 3

4.2.4 Comparison among different Strategies:

Looking at the following Table 3, we can compare the obtained error values for the 3 proposed strategies.

Table 3 - Forecasting errors using different strategies

	Mae [kWh]			Rmse [kWh]			Mbe [kWh]		
Substations	**Strategy 1**	**Strategy 2**	**Strategy 3**	**Strategy 1**	**Strategy 2**	**Strategy 3**	**Strategy 1**	**Strategy 2**	**Strategy 3**
Sub. 1	34.16	19.85	22,61	45.74	28.29	35,51	32.88	-13.65	10,877
Sub. 2	48.98	44.96	48,93	69.87	66.36	73,18	6.27	-3.95	5,579
Sub. 3	10.17	10.09	11,20	16.60	15.07	16,66	7.48	-5.09	8,969
Sub. 4	12.81	10.18	6,59	18.35	14.67	10,47	-0.16	-7.43	0,818
Sub. 5	112.15	56.87	55,39	144.11	77.34	78,85	109.70	-20.29	-23,643
Sub. 6	14.59	9.56	9,65	26.77	20.17	20,21	12.56	-6.04	-5,037
Sub. 7	12.70	7.67	7,20	19.15	11.50	10,46	11.37	-2.58	1,615
Sub. 8	6.61	4.54	7,53	10.09	6.75	12,53	3.82	-1.08	4,210
Sub. 9	21.24	11.74	6,48	28.41	16.62	9,77	21.03	-10.99	-0,052
Sub. 10	92.92	192.92	121,26	177.95	318.38	225,41	-26.72	117.58	-74,154
Sub. 11	73.49	74.37	67,89	125.73	134.20	126,19	27.17	33.32	16,639
Sub. 12	108.97	57.70	14,93	135.24	68.37	23,53	108.97	57.53	7,728
Sub. 13	4.53	3.43	2,39	6.62	4.93	3,54	3.46	-1.71	0,677
Sub. 14	3.49	5.34	2,52	6.72	6.84	3,75	-1.96	4.56	0,831
Sub. 15	10.96	6.11	4,48	16.74	10.09	8,12	-9.99	3.83	-0,151
Sub. 16	4.85	4.88	4,63	11.13	8.33	9,33	-1.83	2.78	0,447
Sub. 17	8.31	5.31	3,57	13.51	10.00	5,84	3.86	-1.04	-1,210
Sub. 18	20.02	16.01	17,62	28.06	25.28	27,16	-6.99	4.89	-2,682
Sub. 19	27.67	5.58	4,47	34.97	8.88	7,71	27.67	-2.58	1,842
Sub. 20	6.20	5.50	4,62	11.18	8.57	7,37	-1.40	-0.48	-1,258
Sub. 21	24.94	10.23	7,50	32.78	15.13	12,87	23.87	0.17	2,680
Sub. 22	5.70	3.64	4,02	9.59	6.04	6,70	-4.35	-1.63	0,667
Sub. 23	9.23	7.34	6,85	14.77	11.53	13,55	-3.87	0.24	-0,090
Sub. 24	11.22	14.13	13,93	20.53	25.11	27,40	-0.41	-9.40	4,596
Sub. 25	12.35	4.76	3,75	15.74	6.77	6,04	11.96	4.17	0,262
Sub. 26	8.21	3.88	2,86	10.64	5.98	4,78	8.12	3.33	0,493
Sub. 27	5.73	5.42	2,68	8.20	7.64	4,45	5.54	5.26	0,238
Sub. 28	6.39	6.14	3,33	10.18	8.56	5,46	0.90	3.81	-0,848
Sub. 29	42.06	12.76	12,47	44.83	16.79	16,88	42.00	-6.93	2,946
Sub. 30	30.74	22.30	24,12	48.78	38.83	63,84	-16.10	-3.96	1,525
Sub. 31	8.53	5.88	7,14	14.07	9.09	11,83	-4.32	2.10	3,528
Sub. 32	5.92	15.96	3,83	9.78	20.03	5,69	1.35	15.77	1,013
Sub. 33	10.47	4.30	3,85	14.45	6.72	5,59	9.86	-2.17	0,995

Looking at the results in the Table 3, we can see that the best strategy is strategy 3, which is the one that uses a neural network for each substation which is then trained on the history of the individual plant. Also looking at the comparison with errors, the strategy turns out to be the one that makes the lowest errors in almost all plants. Another notable thing is that strategy 2 commits slightly higher and comparable errors than the methodology that based on a network for each substation. It is therefore evident that if the number of data available were to increase such methodology based on the identification of clusters and then adopting a

neural network for each cluster identified, it becomes an excellent tool capable of reducing the computational cost that would be had with methodology 3 and with results far more accurate than strategy 2. Specifically compared to strategy 1 we have that strategy 2 results in a 18% reduction in mae, 14% percent for rmse, and 61% percent for mbe. Further, by applying methodology 3 instead of 1 we have a 36% reduction for mae, 25% for rmse and 93% for mbe.

Comparing methodologies 2 and 3 always results in a reduction in percental error in favour of strategy 3, but with lower values. In detail we have that strategy 3 results in a reduction of 22% for mae, 13% for rmse and 81% for mbe. (Table 4)

Table 4 - Average differeces in forecasting strategy

	MAE			**RMSE**			**MBE**		
	Strategy:			**Strategy:**			**Strategy:**		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Mean [kWh]	24,7	20,3	15,8	36,4	31,5	27,3	12,2	4,8	-0,9
Δ% 1	-	18%	36%	-	14%	25%	-	61%	93%
Δ% 2	-	-	22%	-	-	13%	-	-	81%

5. Conclusions and future developments:

In this paper, a method was proposed that succeeds in improving the management of a large amount of data that comes with the management of thermal utilities. The following topics were covered: Data pre-processing, correlation between variables, clustering, and forecasting. It is not possible to identify unique best solutions for thermal clustering and forecast related to DH. This is because DH's controlling systems may be characterized by a variety of configurations, depending on network topology, distribution of energy density demand, type of connected plants, control strategy, environmental conditions etc. [1].

In terms of the case study analysed, the best solution in terms of clustering was obtained through hierarchical clustering, while the best prediction technique was using a neural network for each substation. This result can be attributed to the fact that the number of plants was not so large as to make it computationally difficult to use a single neural network individually. In contrast, if the number of facilities increases, it becomes necessary to adopt the solution that considers one network per cluster. Therefore, research activities should be conducted by increasing the number of samples and substations so that the validity and strength of the method can be increased. The analysis revealed several areas that should be further investigated in the development of an optimal future solution: one improvement would be to create an algebraic combination of the information by relating the plant response, so that the quality of the information to be used to perform clustering could be hardened. As for the forecast, the use of a roll-out strategy could validate this method even more. The installation of additional meters that measure other weather variables or make available access to information regarding building crowding would provide additional exogenous parameters that could be investigated in order to validate both the clustering and forecasting phases.

Ringraziamenti

Innanzitutto, vorrei ringraziare il relatore Prof. Emanuele Ogliari e correlatore Eng. Alfredo Nespoli, la cui continua assistenza e disponibilità durante la realizzazione della tesi è stata fondamentale. La nostra collaborazione mi ha senza dubbio arricchito sia dal punto di vista accademico che umano.

Vorrei inoltre ringraziare sia Michele Puglisi, che mi ha permesso di poter effettuare questo approfondimento mettendomi a disposizione tutti gli strumenti necessari, che Andrea Vieri, il quale mi ha fornito tutto l'aiuto di cui avevo bisogno, sia dal punto di vista organizzativo che concettuale, ma anche un lavoro solido di base che aveva precedentemente impostato su queste tematiche. Grazie a loro e anche a Siram Veolia, poiché conoscere il mondo del lavoro attraverso questa azienda è stato oltre che un onore, un privilegio e sono sicuro che questa esperienza sarà preziosa per il mio futuro.

Ringrazio poi i miei amici di sempre che ci sono sempre stati e sempre ci saranno, la distanza ha provato ad allontanarci, ma ne è uscito un rapporto più profondo e vero.

Mi sento di ringraziare tutti i miei amici del "Tutorato" per aver reso davvero

Speciali questi anni di università.

Dalle prime lezioni fino ad oggi, ogni momento lo porterò sempre con me.

Raggiunto questo obiettivo, spero di condividere con voi altrettanti successi.

Un ringraziamento speciale va alla mia fidanzata Martina che in questi anni è stata sempre al mio fianco, sia nei momenti più facili che in quelli più difficili. Non ha mai smesso di credere in me e nel lavoro che facevo e questo è stato oltre che prezioso anche fondamentale per farmi crescere e imparare a superare gli ostacoli.

Infine, ringrazio la mia famiglia, mia Mamma, mio Papà, mio fratello, i miei Nonni tutti, mio cugino e mia zia. Quattro righe nero su bianco sono solo una minima parte del lungo libro che dovrei scrivere per ringraziarvi. Voglio solo dirvi che grazie a voi e al vostro sostegno e amore sono diventato quello che sono e ho potuto raggiungere tutte gli obiettivi che mi ero posto.

References

[1] S. D. V. V. Elisa Guelpa, «Thermal request optimization in district heating networks using a clustering approach,» *Applied Energy,* pp. 608-617, 2018.

[2] Z. T. Z. M. G. L. Y. L. J. N. Qiang Zhang, "Development of the heating load prediction model for the residential building of district heating based on model calibration," *Energy,* vol. 205, 2020.

[3] S. W. Henrik Gadd, "Heat load patterns in district heating substations," *Applied Energy,* vol. 108, pp. 176-183, 2013.

[4] S. W. Henrik Gadd, "Achieving low return temperatures from district heating substations," *Applied Energy,* vol. 136, pp. 59-67, 2014,.

[5] Z. T. P. P. J. N. W. L. H. Z. Yakai Lu, "GMM clustering for heating load patterns in-depth identification and prediction model accuracy improvement of district heating system," *Energy and Buildings,* vol. 190, pp. 49-60, 2019.

[6] A. Stefan, "Heat load of buildings supplied by district heating. An analysis based on measurements in 50 buildings; Fjaerrvaermekunders vaerme- och effektbehov. Analys baserad paa maetresultat fraan femtio byggnader," Sweden, 1996.

[7] P. D. B Bøhm, "Monitoring the energy consumption in a district heated apartment building in Copenhagen, with specific interest in the thermodynamic performance," *Energy and Buildings,* vol. 36, no. 3, pp. 229-236, 2004.

[8] E. Dotzauer, "Simple model for prediction of loads in district-heating systems," *Applied Energy ,* vol. 73 , p. 277–284, 2002.

[9] C. E. P. a. R. C. S. H. S. Hippert, "Neural networks for short-term load forecasting: a review and evaluation," *IEEE Transactions on Power Systems,* vol. 16, no. 1, pp. 44-55, 2001.

[10] S. S. Haykin, Neural networks : a comprehensive foundation, N.J.: Upper Saddle River, N.J. : Prentice Hall, 1999.

[11] B. E. P. M. Y. H. Guoqiang Zhang, "Forecasting with artificial neural networks: The state of the art," *International Journal of Forecasting,* vol. 14, pp. 35-62, 1998.

[12] D. W. T. H. R. T. Gareth James, An Introduction to Statistical Learning, New York, NY: Springer New York, NY, 2021.

[13] Z. T. M. T. W. C. Xiang Deng, "A clustering-based climatic zoning method for office buildings in China," *Journal of Building Engineering,* vol. 42, 2021.

[14] R. T. J. F. Trevor Hastie, The Elements of Statistical Learning, New York: Springer New York, NY, 2009.

[15] "Matlab," Matlab, [Online]. Available: https://www.mathworks.com/help/stats/kmeans.html. [Accessed 23 Giugno 2022].

[16] "Matlab," Matlab, [Online]. Available: https://www.mathworks.com/help/stats/hierarchical-clustering.html. [Accessed 23 Giugno 2022].

[17] H.-P. K. J. S. X. X. Martin Ester, "A Density-Based Algorithm for Discovering Clusters," *kdd,* vol. 96, no. 34, pp. 226-231, 1996.

[18] P. J. ROUSSEEUW, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987.

[19] S. T. D. Kumar, "Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC)," *Journal of Big Data,* vol. 3, no. 13 , 2016.

[20] Matlab, "Help Matlab Mathworks," Matlab , [Online]. Available: https://www.mathworks.com/help/stats/dbscan.html. [Accessed 25 Giugno 2022].

[21] M. System, "Osservatorio Meereologico di Chivasso," [Online]. Available: http://www.meteosystem.com/dati/chivasso/index.php. [Accessed 25 Giugno 2022].

[22] S. P., "Pearson's correlation coefficient," *British Medical Jornal,* vol. 345, 2012.

[23] Matlab, "Mathworks - Help Center - corrcoef," Matlab, [Online]. Available: https://www.mathworks.com/help/matlab/ref/corrcoef.html. [Accessed 26 Giugno 2022].

[24] E. S. Hub, "European Science Hub," [Online]. Available: https://re.jrc.ec.europa.eu/pvg_tools/en/. [Accessed 25 Giugno 2022].

Abstract in italiano

Advisor:

Prof. Emanuele Giovanni Carlo Ogliari

Co-advisors:

Eng. Alfredo Nespoli

Academic year:

2022-2023

Abstract: Il crescente interesse per l'efficienza energetica e l'onerosa quantità di dati coinvolti rende necessaria l'adozione di un metodo che miri a migliorare la gestione di una serie di edifici collegati a un'utenza termica. Questo lavoro di tesi si propone di fornire un metodo basato sul clustering non supervisionato per classificare le utenze e successivamente utilizzare le reti neurali per prevedere i consumi. Le variabili raccolte dagli smart meters sono alla base dell'analisi, insieme ai dati meteorologici. Il metodo proposto è basato su una vera e propria rete di teleriscaldamento funzionante e quindi vengono utilizzati dati del tutto reali, che quindi sono affetti da buchi e mancate letture, per cui viene eseguita una di pulizia dei dati preliminare e successivamente una post-elaborazione dei dati. L'obiettivo del lavoro è quello di prevedere i consumi attraverso metodi di machine learning come le reti neurali. È stata effettuata una ricerca tra 3 strategie di clustering: k-means, clustering gerarchico e DBSCAN, per stabilire quale sia la più appropriata per il seguente caso di studio. Il clustering gerarchico è risultato essere il più affidabile e con i risultati più convincenti, sulla base degli indici utilizzati per valutare la bontà del raggruppamento. Infine, per poter prevedere il consumo di energia termica richiesto, sono state adottate tre diverse strategie: (1) l'addestramento di una rete neurale per tutte le utenze, (2) l'utilizzo di una rete neurale per ogni cluster e infine (3) l'adozione di una rete neurale per ogni utenza. Analizzando i risultati risulta evidente come la metodologia 3 sia quella che ha portato a risultati migliori, tuttavia non molto distanti dalla strategia 2, la quale potrebbe rivelarsi vincente in caso di aumento dei dati e delle utenze analizzate

Key-words: Machine Learning, Metodi di clustering, Previsione del giorno prima con Reti Neurali, Rete Teleriscaldamento