POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

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

TESI MAGISTRALE IN ENERGY ENGINEERING – INGEGNERIA ENERGETICA

AUTHOR: RICCARDO SCORSOLINI

ADVISOR: PROF. EMANUELE GIOVANNI CARLO OGLIARI

CO-ADVISOR: ENG. ALFREDO NESPOLI

ACADEMIC YEAR: 2022-2023

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 coming from the management of substations of a district heating plant. 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 necessity of the work arises from the fact that the use of full physical models is impossible because most data required for complete characterization of the buildings are unavailable [1]. In addition, the classification, and the forecast can help heating operators meet heat demand in advance and thus formulate wise operation strategies.

2. Proposed Methods

The work begins by collecting measured data from substations, after which a data cleaning is carried out. This step is necessary since the data come from real smart meters and therefore may be affected by missing data. 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. To clustering, the exogenous characteristics that best correlate with energy demand will be selected.
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. [2]
The clustering task was performed by 3 different methods: k-means clustering, hierarchical clustering and the dbscan method. 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 graphical 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.
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. 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. The error calculation and subsequent evaluation is necessary to compare the 3 strategies investigated for prediction. Indeed, it was investigated which was the best forecasting methodology between: using one neural network for all substations; one neural network for each cluster; and finally, one neural network per substation. After identifying the winning strategy, post-processing of the data is performed so that there is effectively zero thermal energy demand at times when the plant is formally off.

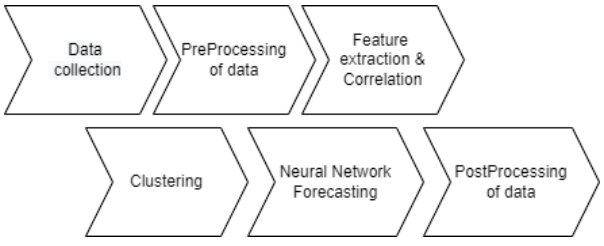

Figure 1 - Flowchart of Methodology

.

3. Case study

The analysis will be conducted based on the data made available by smart meters and those from the appropriate weather station. Veolia's *Esight* platform was used to collect the data from the meters. 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 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.
The next stage of data collection is carried out by collecting available data from the meteorological station at the site of interest [3]. Since the telecontrol network has intermittent operation, indeed it operates only at certain times of the day and year an investigation was conducted, whether it would be worthwhile to use a parallel dataset that considers only actual readings of smart meters excluding zeros due to plant shutdown.

3.2 Feature Extraction

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. Correlating the required thermal energy with all other variables we have, what is summarized in the table.

Table 1 - Correlation Coefficient between Energy required a and other variables

	Mean Corr. Coef. T ON	Mean Corr. Coef. 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 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 period. The same was done with weather variables to observe what level of correlation one had with thermal energy for prediction purposes, indeed these variables were all used as input parameters for the clustering phase.

3.3 Clustering

The clustering activity is carried out by seeking which parameters correlate best with thermal energy. 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.*

Before implementing the clustering techniques, an investigation of the presence of outliers that would affect the result and the correct representation of them was carried out, for this reason a boxplot was developed to observe the presence or absence of any outliers. As we can see from Figure 2, 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 3) 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 stage of clustering and treated precisely as outliers.

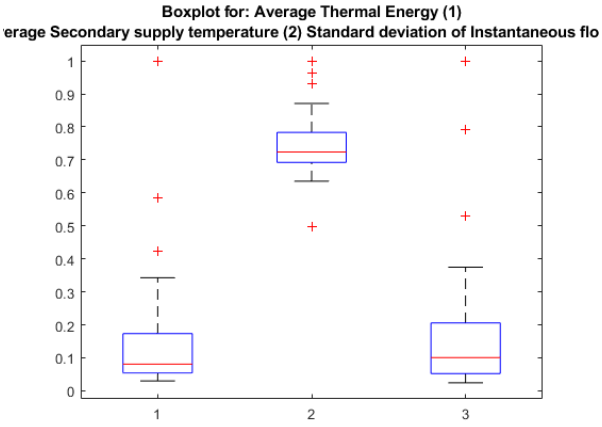

Figure 2 - Boxplot

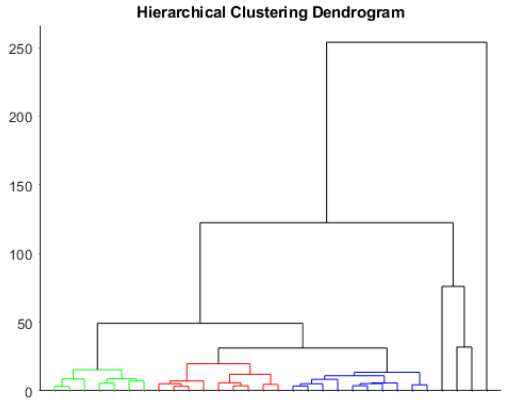

Figure 3- Hierarchical Dendrogram with outlier

3.4 Forecasting

After clustering the data, the method considered proposes 3 strategies so that we can investigate which one best fits the case study. 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 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. 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. 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 to the neural network. 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.

3.3.1 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.

4. Results

4.1 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.

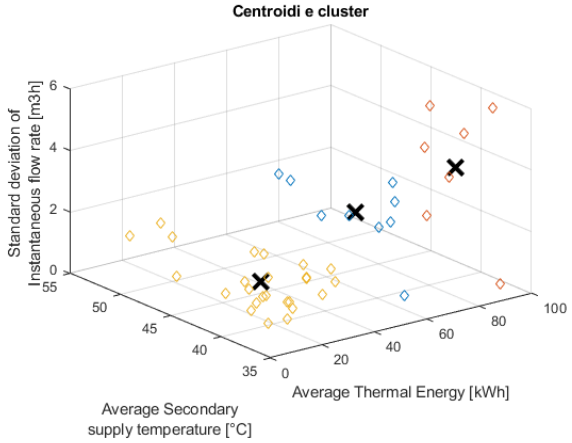

Figure 4 - k-means Clustering Spatial Representation

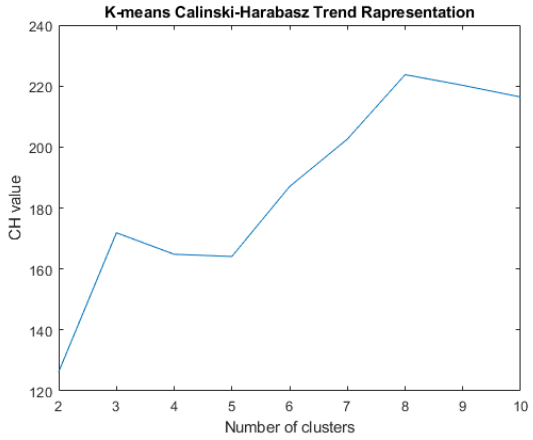

Figure 5 - K-means Calinski-Harabasz Trend Rapresentation

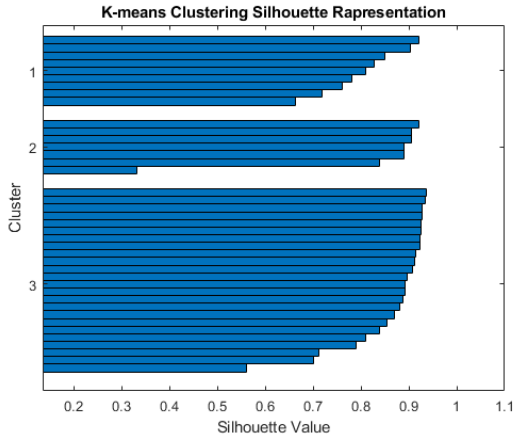

Figure 6 - k-means Silhouette Representation

Looking Figure 4 at we can see that the clustering is pretty omogeneous, confirmed by the Silhouette index (Figure 6) which show that the distinction between the clusters is quite clear. In addition, Figure 5 shows that the number of groups is consistent with the CH trend indeed we have a

negative inflection point around the 3 clusters. To further investigate the most populous cluster subdivision, the other techniques listed below are adopted.

4.2 Hierarchical Clustering Results

Using the same variables adopt previously and applying hierarchical clustering, with a cut-off value of 30, we obtain that the cophenetic index is: 0,8334. A value very close to 1, which proves the goodness of clustering.

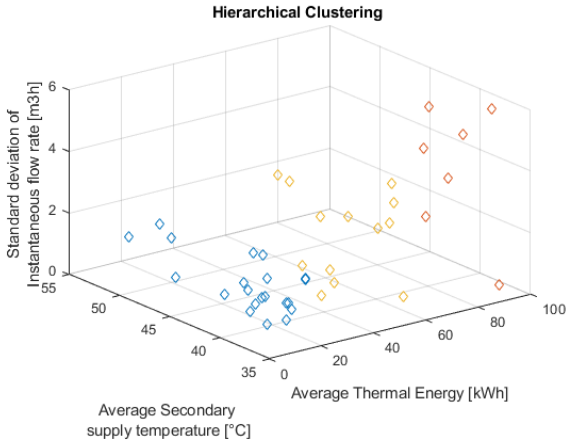

Figure 7 - *Hierarchical Clustering Spatial Representation*

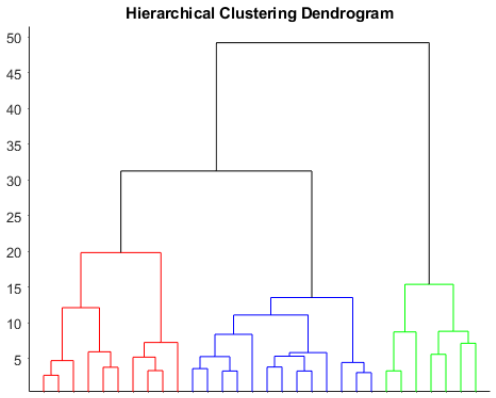

Figure 8 -Hierarchical Dendrogram

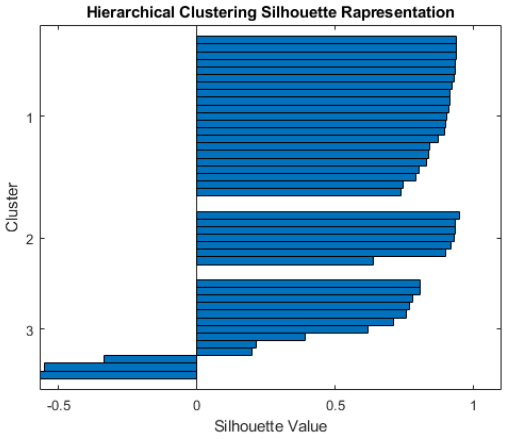

Figure 9 - Hierarchical Clustering Silhouette Representation

As we can appreciate from both the dendrogram (Figure 7) and the spatial arrangement (Figure 8), the arrangement of the three clusters is clear. 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, which appear uniform, except for three items found in cluster 3. Therefore, an analysis is attempted using the DBSCAN technique, which is suitable for breaking up various clusters [4] [5].

4.3 DB-SCAN Clustering Results

Before applying the DBSCAN algorithm, a preliminary analysis is performed to determine the best value of MINPITS and EPSILON. 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. To find the value of epsilon instead, we start from the k-graph at its knee point (figure 10). On subsequent analyses, however, this value was reduced to unpack large clusters, hence the optimal value chosen for this case study is 9. [5]

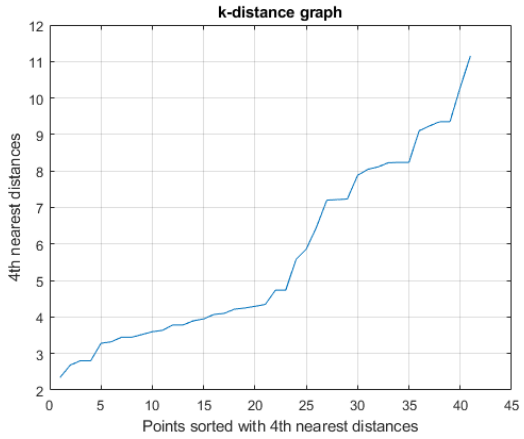

Figure 10 - DBSCAN k-distance graph

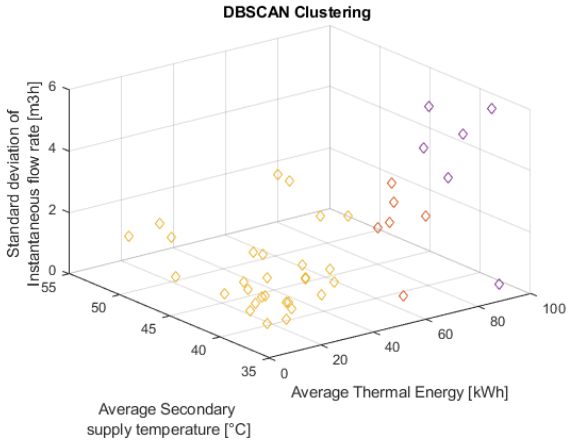

Figure 11 - DBSCAN Spacial Representation

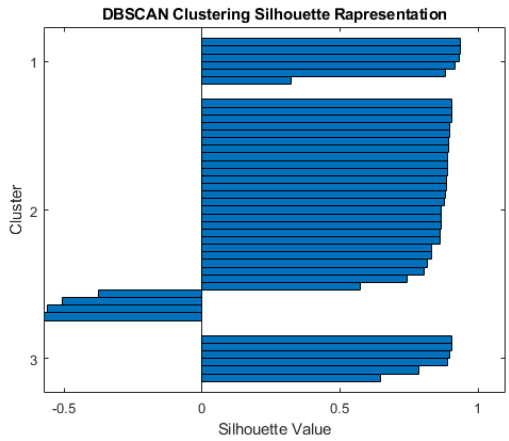

Figure 12 - DBSCAN Silhouette Index

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 Index (Figure 12), indices are affected by negative values. This representation confirms what has already been obtained in hierarchical clustering.

4.4 Clustering Results

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. 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 [1]

4.5 Neural Network Post-Processing

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.

Table 3 - Forecasting errors using different strategies

	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%

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

[1] 4 utilities were eliminated from the analysis because the available data were too minimal to be investigated

then trained on the history of the individual plant. 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.

5. Conclusions

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]

Regarding this 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.

Bibliografia

[1] V. V. Elisa Guelpa, «Thermal request optimization in district heating networks using a clustering approach,» *Applied Energy*, vol. 252, pp. 608-617, 2018.

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

[3] M. System, «Osservatorio Meereologico di Chivasso,» [Online]. Available: http://www.meteosystem.com/dati/chivasso/index.php. [Consultato il giorno 25 Giugno 2022].

[4] H.-P. K. J. S. X. X. Martin Ester, «A Density-Based Algorithm for Discovering Clusters,» *kdd*, vol. 96, n. 34, pp. 226-231, 1996.

[5] Matlab, «Help Matlab Mathworks,» Matlab , [Online]. Available: https://www.mathworks.com/help/stats/dbscan.html. [Consultato il giorno 25 Giugno 2022].