Nadezda Belonogova, Markku Pulli

Applying machine learning to cluster residential electricity customers based on their load profiles and outdoor temperature

Vaasa 2021

Contents

1 Introduction to the topic

This chapter presents the background, motivation and objectives of the study.

1.1 Background and motivation

The electricity consumption patterns are changing with the electrification of heating sector, contributing to an increase of weather-dependent load in residential building stock. This, in turn, represents a significant flexibility potential that remains untapped mainly due to uncertainty and lack of regulatory framework for both consumers and energy stakeholders to provide and access flexibility resources, respectively.

At the same time, digitalization of power systems together with developing ICT services are enabling numerous benefits such as smart meter roll-out, smart charging, demand-side flexibility activation and verification. The major motivation drivers for developing clustering methods for electricity consumption are following:

- digitalization of power systems, smart meter roll-out
- Big data era, progress in development of data analytics tools
- field trials to quantify demand-side flexibility, for small datasets and high temporal granularity
- energy stakeholders(distribution system operators, electricity retailers, aggregators) still lack tools to benefit of smart meter data and integrate it into their operating and planning practices
- digitilization, emergence of datahubs
- decarbonisation goals, need for demand-side flexibility in power system
- socio-demographic information is limited due to privacy issues

The smart data analytics has been given a considerable attention during the past several years (Westermann, Deb, Schlueter, and Evins (2020), Niu, Wu, Liu, Huang, and Nielsen (2021)).

We have already seen numerous studies on clustering using sub-hourly resolution data and a limited number of customers (Rajabi et al. (2019), Quilumba, Lee, Huang, Wang, and Szabados (2015)), synthetically modelled load profiles Fischer, Surmann, Biener, and Selinger-Lutz (2020) at the sub-hourly resolution.

In Ofetotse, Essah, and Yao (2021), additional data such as socio-economic, building or appliance factors were used together with electricity consumption data to provide better clustering results.

Wang, Chen, Kang, and Xia (2016) implemented a time-based Markov model from the consumption dataset and used it as a feature for the clustering technique Fast Search and Find of Density Peaks (CFSFDP). In Wen, Zhou, and Yang (2019), Kmeans clustering method was used together with PCA for the time-series consumption dataset.

In Cerquitelli, Chicco, Corso, et al. (2018) and Cerquitelli, Chicco, Di Corso, et al. (2018), Kmeans and dynamic time warping clustering techniques were used for hourly load datasets. One of the highlights of the research work was, that there is no common approach for clustering consumption data. Instead, selection of relevant features and parameters should be data-driven.

This brings us to the objectives of the project work.

1.2 Objectives and scope

The main objective of the work is to find the most suitable machine learning architechture to distinguish in the given dataset various heating systems. This includes developing,

testing and evaluating various clustering techniques together with data preprocessing, dimensionality reduction and feature selection techniques.

The expected results are the meaningful clusters of customers with same heating systems and combinations of heating systems. Description of clusters and identification of heating systems is outside of the scope of the project. Another expected result is identifying relevant features and a suitable machine learning architecture to obtain the clusters.

The presence of various heating systems in the dataset, and in addition to that, various combinations of heating systems on a single customer's premises increases the complexity of the problem. At least the following heating systems are present in the dataset:

- non-electric heating (oil-based or district heating), that is not dependent on outdoor temperature
- only direct electric heating
- direct electric heating together with auxiliary non-electric heating, e.g. woods
- direct electric heating together with auxiliary electric-heating such as heat pumps
- heat pumps plus auxiliary woods heating
- heat pumps with both heating and cooling functionalities
- stored electric heating

2 Machine learning data flow

The data flow in machine learning represents the steps that are needed to reach the set objectives (see Chapter 1.2). These include data preprocessing, feature engineering, model construction and validation and finally, visualization and interpretation of results. These are presented on Figure 1.

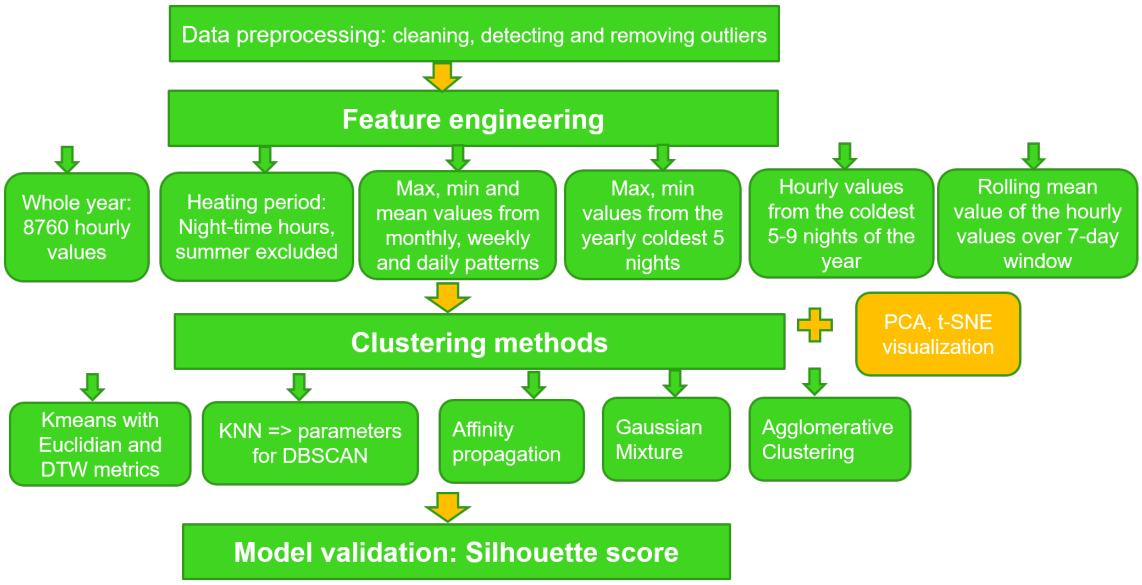

Figure 1. Illustration of machine learning data flow.

As it can be seen from the figure, various feature vectors were built and tested together with various clustering techniques.

The steps will be described and the results of each of them analyzed in more detail in the following subchapters.

2.1 Input data

The data contains hourly electricity consumption measurements for about 50 000 customers during a one-year period, see Figure 2

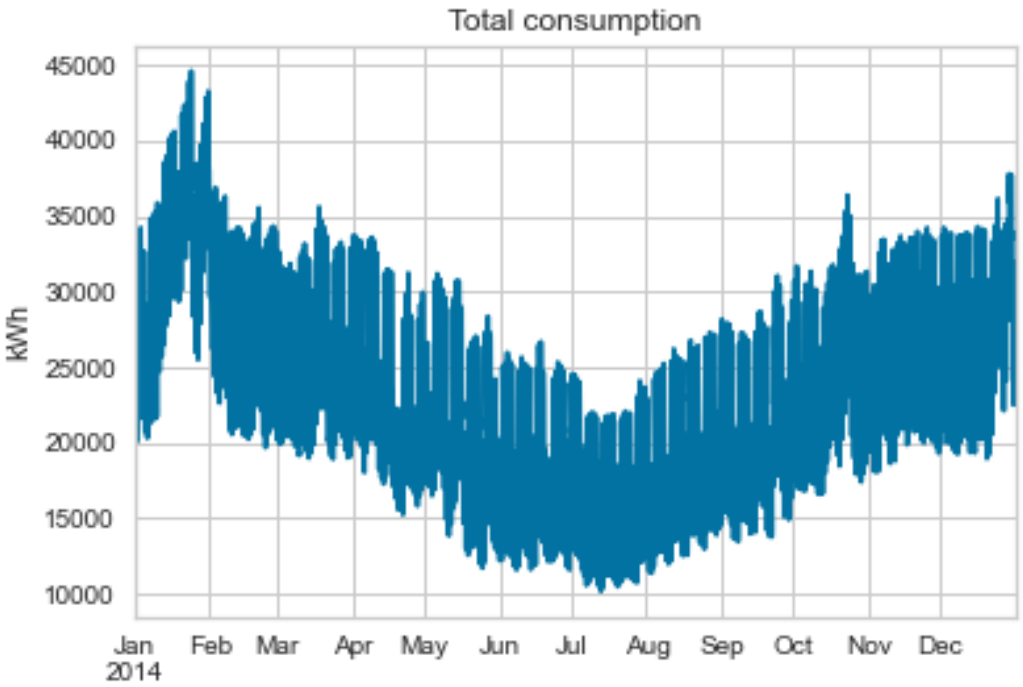

Figure 2. Total electricity consumption of ca.50 000 customers.

The load profile of the total consumption indicates the presence of outdoor temperature-dependent electricity heating loads on the customers' premises. We can see that the load levels are high during colder winter months and lower during summer months.

In addition to that, hourly outdoor temperature is also given for the area where the customers are located, and presented on Figure 3.

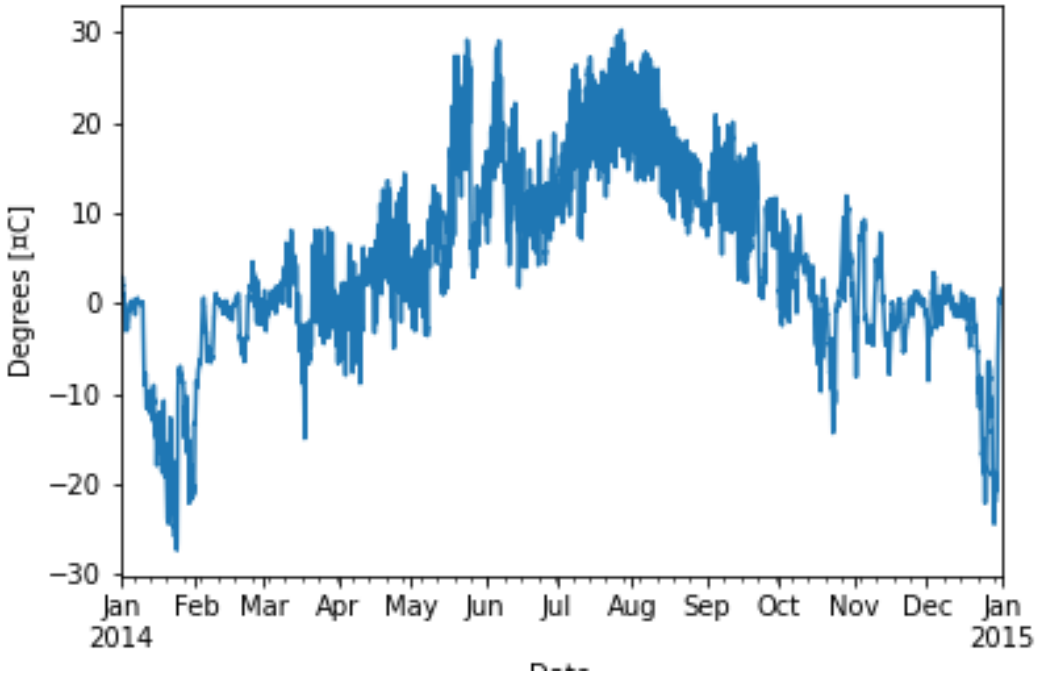

Figure 3. Hourly outdoor temperature.

2.2 Data preprocessing

Before extracting features and clustering process, data preprocessing is an important step in order to eliminate zero or nan measurement values as well as remove the outliers.

The data was cleaned using the following steps:

- customers with missing zero or nan values were removed
- customers with annual consumption larger than 60 000 kWh and less than 1000 kWh were also removed, since most likely they are either not residential customers or summer cottage houses, respectively.
- customers with annual maximum power value higher than 100 kW and minimum

value higher than 10 kW were excluded from some of the analyses.

In the project, original, scaled (MinMaxScaler) and normalized data have been used for clustering. Data scaling was column wise and normalization was done row basis. Only one method was used at a time.

2.3 Feature engineering

During the project work, several approaches for creating a feature vector have been applied and tested in clustering methods.

Depending on the clustering objective(s), the feature vector is created appropriately. In this work, the objective is to distinguish various heating systems and therefore, load pattern is important. Another objective could be for instance, clustering based on size of house, presence of electricity spot market-based load control, solar PV or electric vehicle charging load. In that case, different feature vector should be obtained.

The feature vector should on one side, include all the relevant characteristics of a load profile to reflect the outdoor temperature dependency and by that way, correctly assign a given load profile to the cluster, but on the other side, exclude the irrelevant noise due to daily activities of habitants, cyclic appliances (electric water heater), sauna activities and others that are not dependent on outdoor temperature and hence do not reflect the heating system type.

In the context of this topic, a feature may be electricity consumption value for a single hour, maximum, minimum and average values over a specific period of time, or a mean rolling value over a specified period of time. These options have been implemented in this work. However, other numerous alternatives are possible and worth of experimenting with. These may be, for instance, substitute the original time-series data with the sorted values in descending order during every day,week or month. Another option could be

taking derivative over the time-series profile to reveal the nature (fast, slow) of power changes from hour to hour. Further option could be converting the time-series data into a cumulative profile, where each hour's value is equal to the sum of the previous hours' values.

2.3.1 Feature vector 1

The first approach was taking all 8760 features. However, this was too noisy and also heavy to carry out the clustering. Model training time very high and low performance.

2.3.2 Feature vector 2

The second approach was to select only particular periods of year that are significant in terms of heating. Thus, summer months (May-August) were excluded out of clustering. Furthermore, only night-time hours from 22:00 to 06:00 were selected to eliminate the noise coming from daily activities of people.

Such approach resulted in ca. 2500 features per sample. It produced some meaningful results with the Kmeans clustering algorithm using DTW metrics. However, the model fitting took too much time. After this, paralellization was applied using 20-30 cores, however it turned out that the paralellization of DTW is not a straightforward task and requires more elaboration and usage of additional packages.

2.3.3 Feature vector 3

The third approach was to downsample the dataset into seasonal, monthly, weekly and daily periods and select the mininmum, maximum and average values of electricity consumption into the feature vector (see Figure 4)

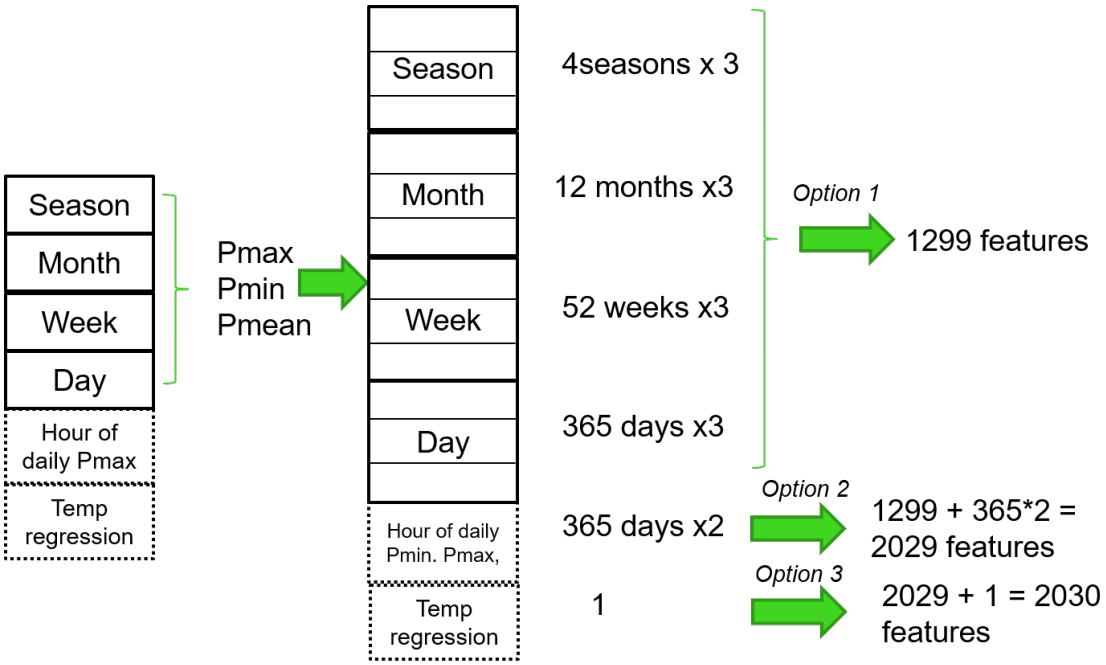

Figure 4. An approach to build a feature vector.

2.3.4 Feature vector 4

The fourth approach was to select the minimum and maximum values over the period of the five coldest nights of the year. This feature vector was short and thus clustering process took little time. This allowed to try several clustering algorithms (see Figure 1). However, none of those provided meaningful results because such short feature vector does not reveal the characteristics of various heating systems.

2.3.5 Feature vector 5

This feature vector was constructed by using the time-series profile of the 7 days of the coldest period of the year. This resulted in 192 features per each sample. Now when the feature vector was relatively short but at the same time it reflects the heating system peculiarities, the clustering results were the best of all the other results obtained with the previous four feature vectors.

2.3.6 Feature vector 6

This feature vector was obtained by averaging the data over the rolling window which width was selected to be 7-day.

Below is the code snippet to get the feature vector for original time-series data and averaged data, as well as outdoor temperature for the coldest period of the year.

```
84  def getColdPeriod():
85      coldTimeLoad1 = data.loc['2014-01-20 00:00:00':'2014-01-24 23:00:00'].T
86      coldTimeLoad2 = data.loc['2014-01-27 00:00:00':'2014-01-29 23:00:00'].T
87      #coldTimeLoad = pd.concat([coldTimeLoad1,coldTimeLoad2],axis=1)
88      coldTimeLoad = coldTimeLoad1
89      return coldTimeLoad
90
91  def getColdPeriodRoll():
92      roll_period = 7
93      coldTimeLoad1 = data.loc['2014-01-20 00:00:00':'2014-01-24 23:00:00']
94      coldTimeLoadRoll = coldTimeLoad1.rolling(roll_period).mean()
95      return coldTimeLoadRoll.iloc[roll_period-1:,:]
96
97  def getColdPeriodTemp():
98      coldTimeTemp = tempDF.loc['2014-01-20 00:00:00':'2014-01-24 23:00:00']
99      return coldTimeTemp
```

Figure 5. Data for model tuning..

2.4 Model construction

Different number of clusters was tested with different clustering methods (see tested methods on Figure 1). The dataset was split into training and testing sets in the ratio of 0.75.

To find out optimal cluster count two different methods were used. Elbow curve and silhouette score. Two were selected to verify results of each method.

The elbow curve is based on Within-Cluster Sum of Square method. This is the sum of squared distance between each point and the centeroid in a cluster. Once the sum ,based on k-value, starts flatten with x-axis this is the elbow point and optimal k value.

The silhouette coefficient (SC) is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. SC range is from -1 to 1. ’1’ represents the best inter cluster separation, close to ’0’ means clusters are overlapping and ’-1’ corresponds to wrong cluster assignments.

Based on the results obtained during the project work, DTW metrics for the provided dataset. This may be explained by the fact, that the electricity consumption of different households may react to the outdoor temperature with various delays depending on numerous factors, such as house insulation levels, heating demand, etc. This means, that even though two electricity consumption time-series are not synchronized in time, they may still belong to the same heating system and thus to the same cluster. Hence, DTW metrics performs better than Euclidian metrics because it can catch those delays.

The challenge is that DTW metrics is a very time-consuming and computationally exhausting algorithm. For instance, clustering of 10 000 samples with feature vector length of 2500 data points took over 8 hours.

After that, shorter feature vectors containing normalized and non-normalized dataset, as well as original and averaged over a rolling window of 7 days, were tested with Kmeans method and DTW metrics. After that, the model was validated using Silhouette scores for both training and testing datasets.

In addition to DTW metrics of Kmeans clustering technique, the DBSCAN clustering was also tested. This algorithm requires two attributes: epsilon and minimum samples. The epsilon parameter range was estimated using K-Nearest Neighbours approach. After that, a range of epsilon and minimum samples were generated and Silhouette score calculated for numerous combinations of those two attributes.

However, it provided very unstable results since the number of clusters varied very drastically with small changes in epsilon and minimum samples. Furthermore, Silhouette score did not reach high enough values. Therefore, DBSCAN was left out of further simulations.

The agglomerative clustering was also tested, but it was again computationally-exhaustive to execute even with the short feature vectors.

From the silhouette coefficient (see Figure 8)and elbow curve (Figure 7) result analyses it was not so clear what the correct cluster count is. Both methods indicated that the optimal number of clusters is from 2 to 4. This is due to time series data being very similar between the customers and different heating systems. Final model tuning was completed focusing on cluster counts 3 to 5. Cluster count 2 was dropped as separating heating systems only to two clusters would not bring much value for this effort. After model output and performance metric analysis, the optimal count was picked. Silhouette coefficient analysis for clusters using DTW method took 18503 seconds in total.

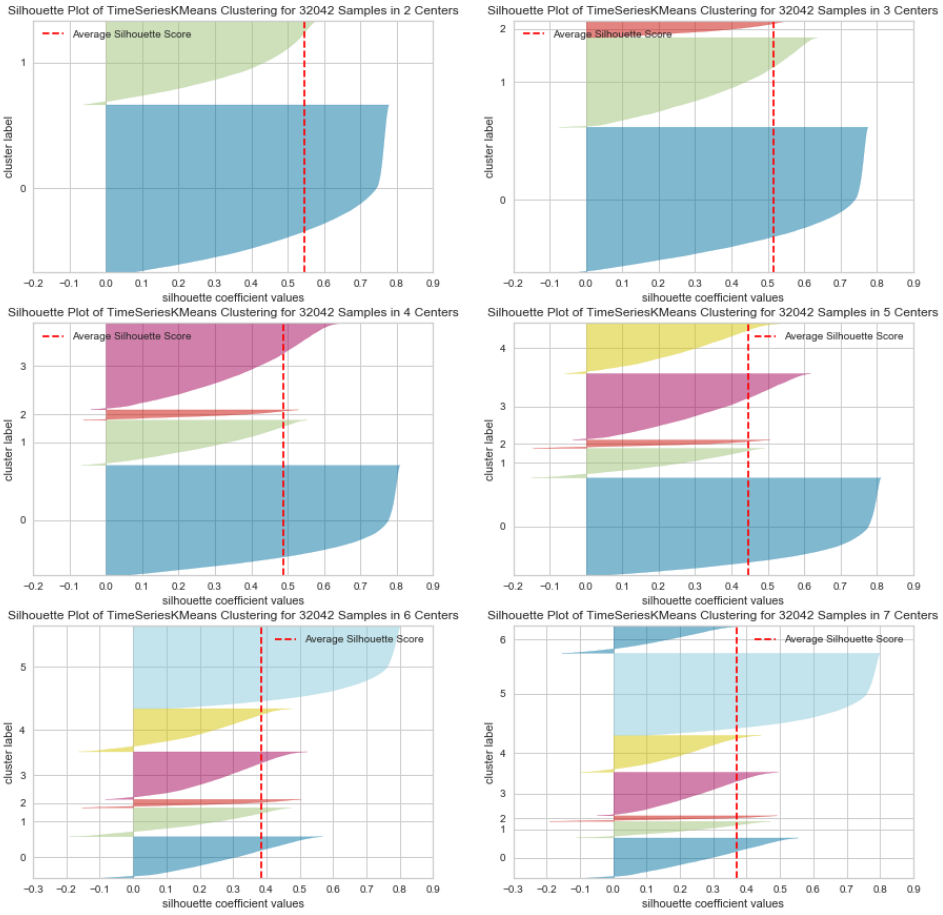

Figure 6. Illustration of silhouette coefficient for k values from 2 to 6..

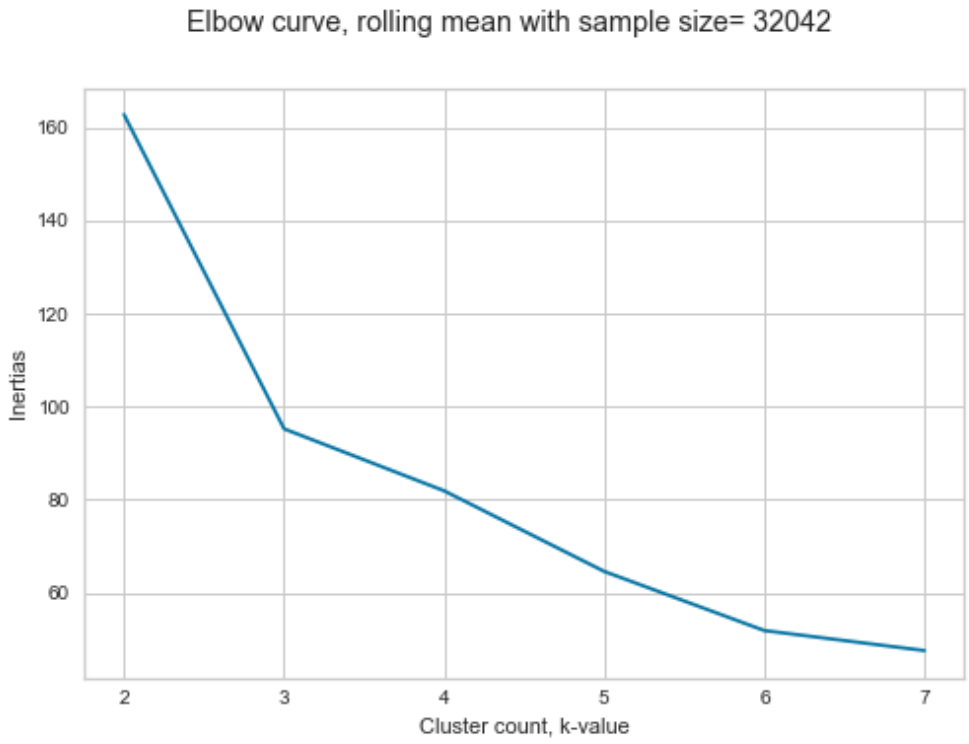

Figure 7. Illustration of the elbow curve. k values from 2 to 7..

Below is the code snippet to build the final time series model for clustering.

```
#Fit the timeseries cluster model
#DTW takes long time to process. From the experience the DTW is more accurate for timeseries clustering than euclidean.
sz = X_train.shape[1]
clusters = 3
seed = 0
np.random.seed(seed)
model = TimeSeriesKMeans(n_clusters= clusters,
                         n_init=1,
                         max_iter=4,
                         metric='dtw',
                         #max_iter_barycenter=2,
                         random_state=24,
                         n_jobs=-1,
                         verbose=True)
%time model.fit(X_train)
#Save the model in pickle format.
filename = '122_Cold_hrs_dtw_model_3C.pkl'
pickle.dump(model, open(filename, 'wb'))
```

Figure 8. Building of the final Kmeans clustering model with DTW metrics..

3 Results and main conclusions

In this section, clustering results obtained with Kmeans method using DTW metrics are presented for the feature vectors built of the annually coldest one-week electricity consumption hourly values.

3.1 Visualization of results

The clustering results are illustrated for number of clusters 3, 4 and 5, for the averaged datasets, all obtained with the Kmeans clustering using DTW metrics. The rest results obtained for the original, scaled and normalized data are presented in Appendix.

3.1.1 3 clusters

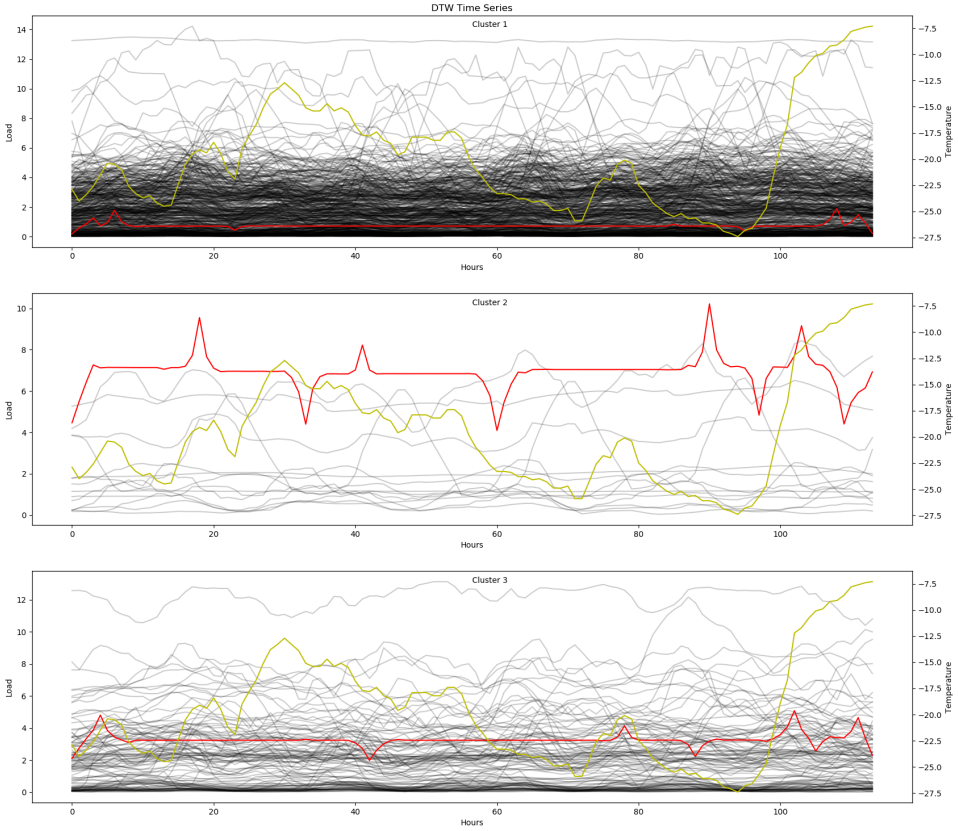

Figure 9. 3 clusters obtained from the averaged dataset.

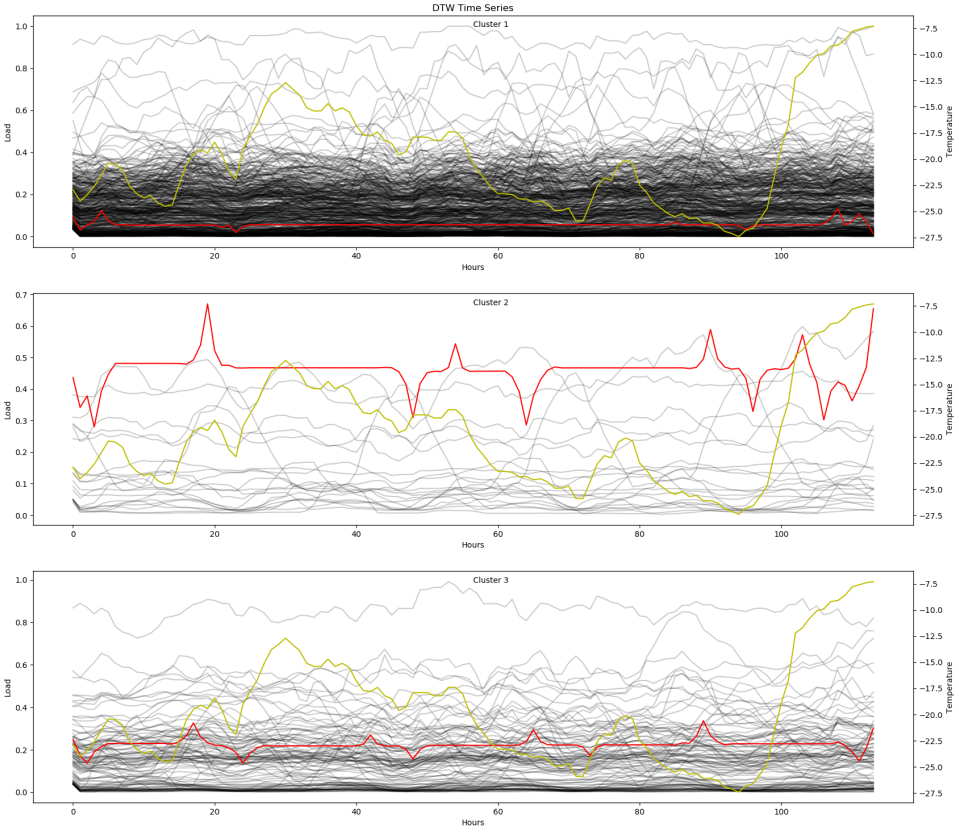

Figure 10. 3 clusters obtained from the averaged and scaled dataset.

3.1.2 4 clusters

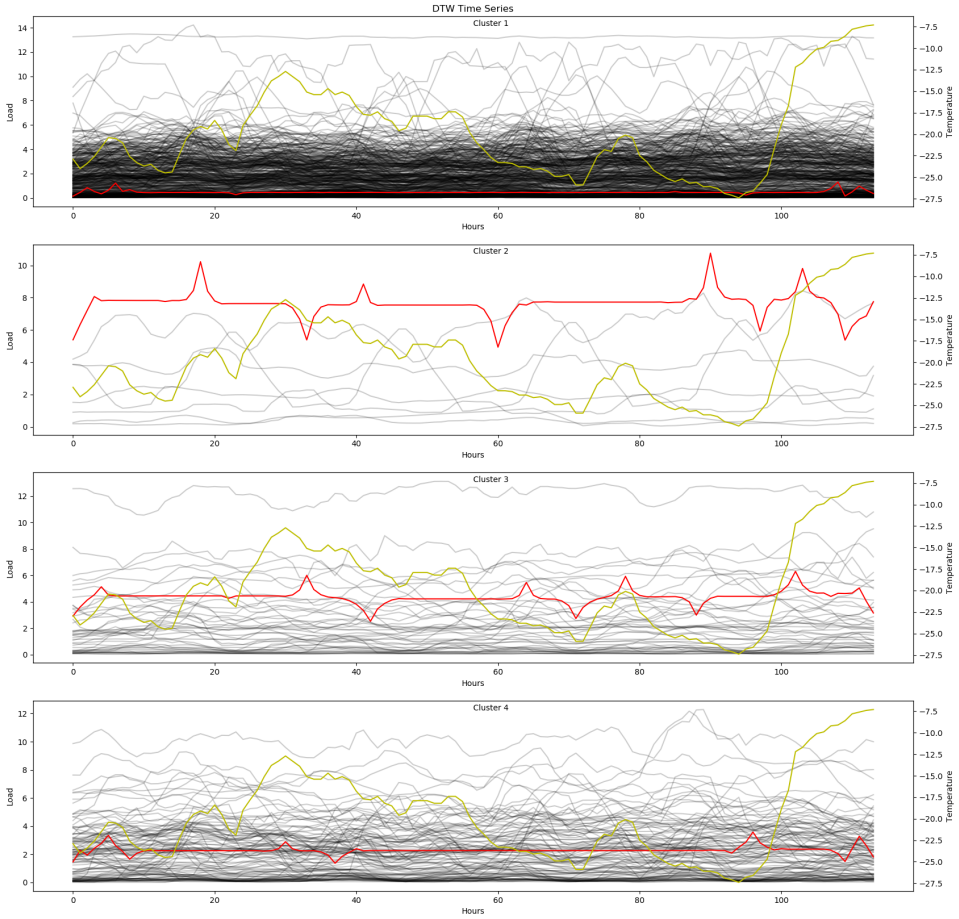

Figure 11. 4 clusters obtained from the averaged dataset.

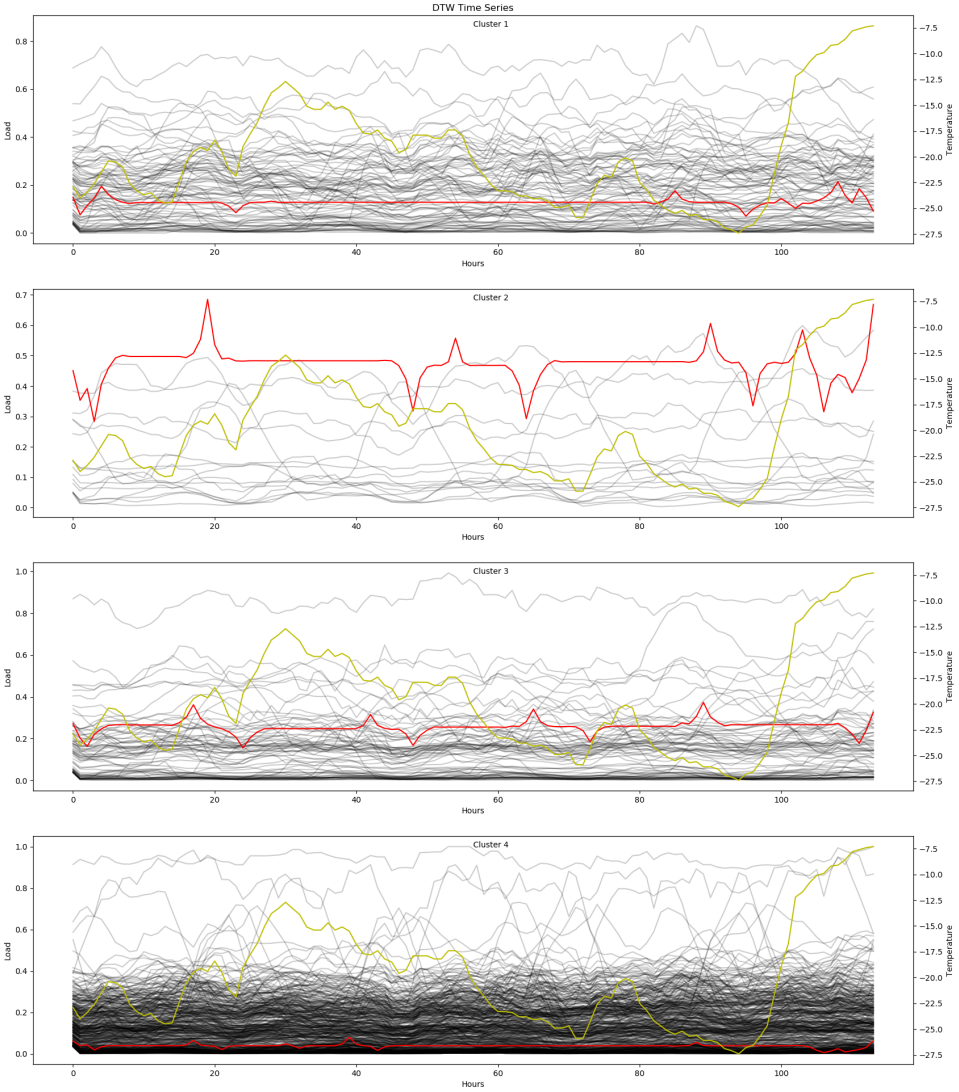

Figure 12. 4 clusters obtained from the averaged and scaled dataset.

3.1.3 5 clusters

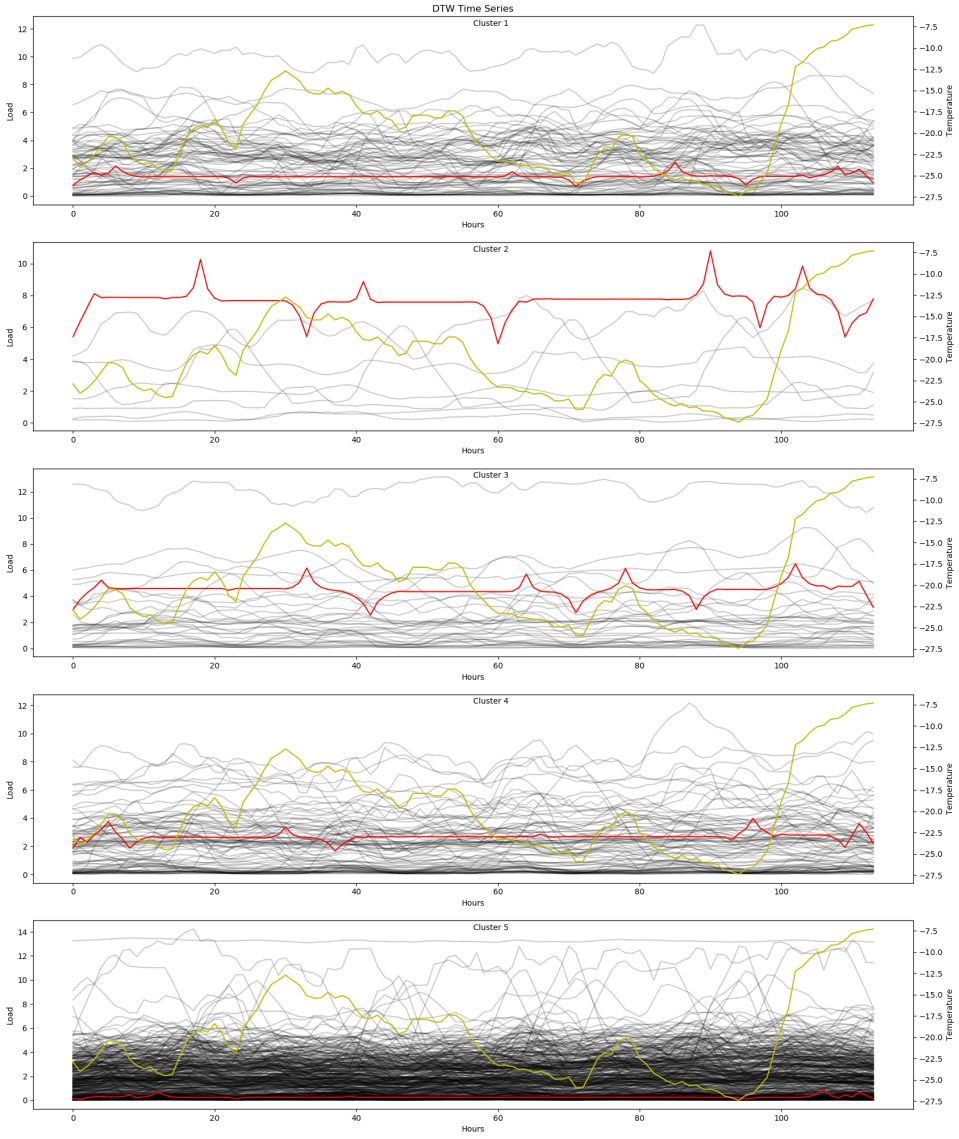

Figure 13. 5 clusters obtained from the averaged dataset.

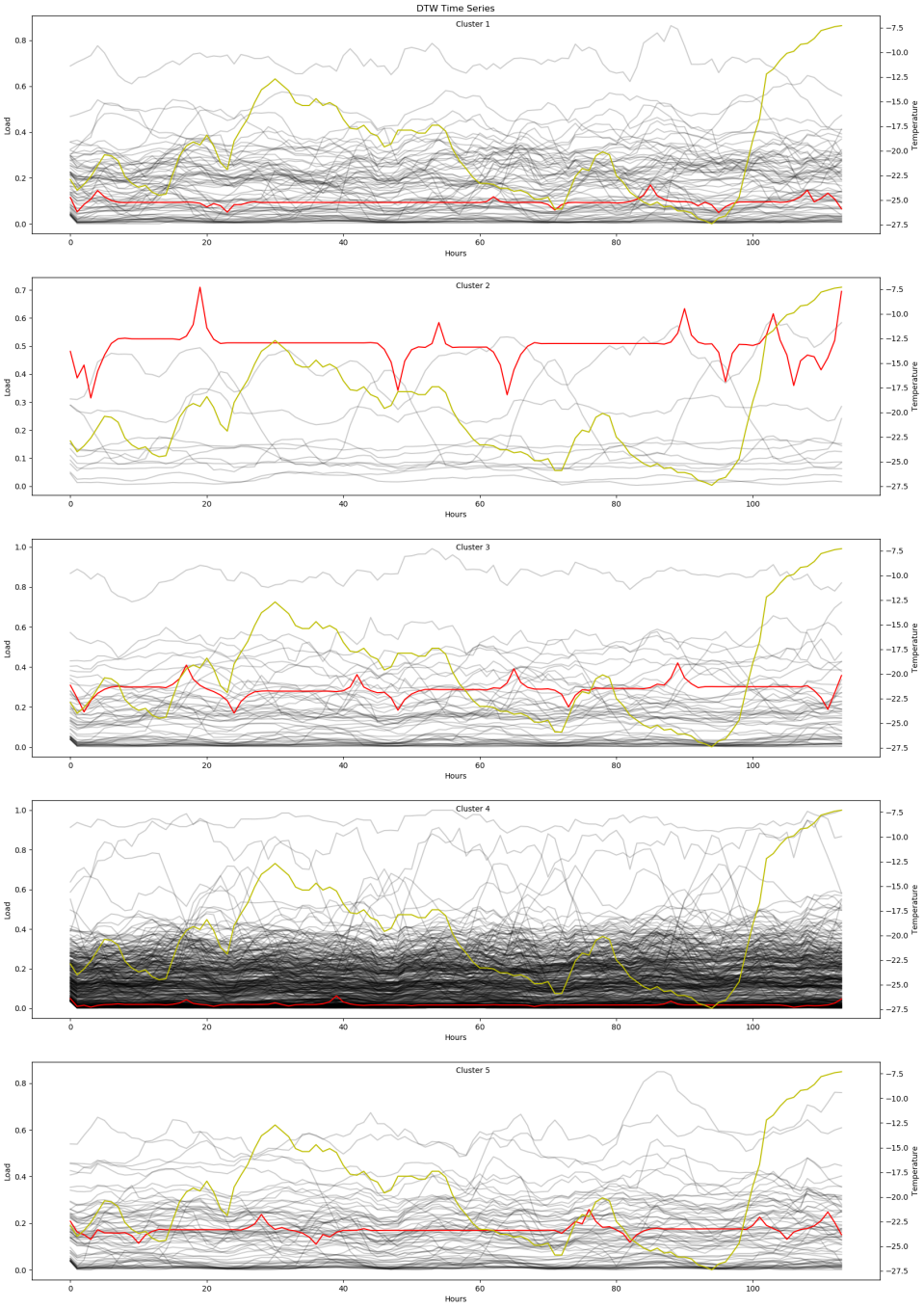

Figure 14. 5 clusters obtained from the averaged and scaled dataset.

3.2 Validation of results

The validation has been done using Silhouette score. The silhouette scores are presented for all feature alternatives in Figure 15.

Raw Hourly			75 %	25 %	Training data				
Cluster count	**Hours**	**Normalizing / Scaling**	**Train silhouette**	**Test Silhouette**	**Cluster0**	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**
3	120	None	0,41146	0,3999	5088	14045	12641		
3	120	Sca	0,43215	0,425	12673	2778	16323		
4	120	None	0,3542	0,3353	6344	12655	11118	1657	
4	120	Sca	0,3912	0,3757	5397	1280	13745	11352	
4	120	Nor	0,0895	0,0192	9781	1882	5217	14894	
5	120	None	0,2638	0,2578	7636	2437	7064	10770	8418
5	120	Sca	0,33439	0,31354	3566	1109	11790	6835	8474
5	120	Nor	-0,0450	-0,0951	8520	3425	6694	2549	10586
Rolling mean									
Cluster count	**Hours**	**Normalizing / Scaling**	**Train silhouette score**	**Test Silhouette score**	**Cluster0**	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**
3	120	None	0,5136	0,5150	17442	2078	12522		
3	120	Sca	0,5117	0,5126	17224	2522	12296		
4	120	None	0,4852	0,4707	13480	1068	5926	11568	
4	120	Sca	0,4646	0,4438	10083	1958	7924	12077	
5	120	None	0,4157	0,3914	7416	1009	4847	8349	10421
5	120	Sca	0,4073	0,3799	6825	1371	5656	10232	7958

Figure 15. Silhouette scores for all calculated scenarios with Kmeans method and DTW metrics.

Based on the results, we can write about the following findings:

- the best clustering results with the highest Silhouette score were obtained using the average dataset as a feature vector. This can be explained by the fact, that the sharp peaks and valleys are smoothed out by applying rolling method making the noise level of the dataset lower.
- the Silhouette score for the training and testing datasets were close to each other, which is good.
- 3 clusters is the optimal number of clusters for the dataset provided that excludes outliers and small customers. This means, that the provided feature engineering and clustering techniques are not capable to identify more combinations of heating

systems. For this, either higher data resolution is required, or further development in feature engineering, as will be described in the next subsection.

- having cluster centroids and data on outdoor temperature, it is still difficult to say what kind of heating systems stands behind each cluster. For this, ground truth values are needed at least for part of the dataset, in order to be able to estimate heating systems of the rest customers.

3.3 Main conclusions and further research needs

The main conclusions from the carried out project work are the following:

- feature engineering turned out to be the most time-consuming part of building machine learning architecture to cluster the electricity consumption data;
- a deeper knowledge of the background of the electricity customers results into better clustering outcomes. For instance, in this work, outlier detection helped to decrease the noise of the data significantly;
- DTW metrics performs significantly better than Euclidian metrics, due to the nature of temperature dependency of electricity consumption;
- the machine learning data flow designed in this work, does not necessarily scales up to other electricity consumption datasets. Feature engineering should be tested and carefully selected separately for each dataset, taking into account its characteristics and all possible background information.

The further research need is sensitivity analysis in respect to feature engineering and metrics of clustering methods. For instance, feature vectors should be constructed also of the other time periods than the one considered in this work (20.1-29.1). Thus longer and shorter time periods should be tried and performance scores (e.g. Silhouette score)

compared between each other. Such analysis will indicate, firstly how stable or unstable results the selected clustering method generates, and secondly, what is the best feature vector for the provided dataset.

Similar approach is a twin sample validation with the difference, that the feature vector length is same but taken from the different time period. For instance, the electricity consumption values taken from one week in January could be clustered and compared to the consumption values taken from one week in February. Having obtained two sets of cluster labels, their similarity can be measured using e.g. F1-measure or Jaccard Similarity. A set of clusters having high similarity with its twin-sample is considered good, which indicates that the clustering is not sensitive to the time period selected.

Another type of sensitivity analysis is finding the best metrics for a clustering algorithm. This may be time-consuming and computationally-exhausting, therefore it should be first tested on a smaller dataset and shorter feature vector.

Further research need is automatization of the outlier detection procedure. For instance, removing customers with 0 or nan values, substituting nan values with certain logic, or deleting too large and too small customers (based on total annual energy and/or maximum annual power) is a straightforward and easy to implement approach. However, there may be further outliers within the selected group of samples. In the case of electricity customers, after the preprocessing procedure described above, there may be still a lot of customers that do not represent residential houses but still fit into the residential customer limits. These are for instance, estate-related consumption in the block of flats or row houses such as outdoor or indoor lighting, parking lots, elevators, ventilation, etc. These should be also removed from the clustering process because the electricity consumption pattern does not represent a single household customer. Furthermore, there may be cases when a group of row apartments have a single electricity automatic meter reading (AMR) device and hence the electricity consumption represent the consumption of a group of houses, but not a single house. This type of customers may also disturb the clustering process.

Bibliography

Cerquitelli, T., Chicco, G., Corso, E. D., Ventura, F., Montesano, G., Armiento, M., ... Santiago, A. V. (2018). Clustering-based assessment of residential consumers from hourly-metered data. In *2018 international conference on smart energy systems and technologies (sest)* (p. 1-6).

Cerquitelli, T., Chicco, G., Di Corso, E., Ventura, F., Montesano, G., Del Pizzo, A., ... Sobrino, E. M. (2018). Discovering electricity consumption over time for residential consumers through cluster analysis. In *2018 international conference on development and application systems (das)* (p. 164-169).

Fischer, D., Surmann, A., Biener, W., & Selinger-Lutz, O. (2020). From residential electric load profiles to flexibility profiles – a stochastic bottom-up approach. *Energy and Buildings*, *224*, 110133. https://doi.org/https://doi.org/10.1016/j.enbuild.2020.110133

Niu, Z., Wu, J., Liu, X., Huang, L., & Nielsen, P. S. (2021). Understanding energy demand behaviors through spatio-temporal smart meter data analysis. *Energy*, *226*, 120493. https://doi.org/https://doi.org/10.1016/j.energy.2021.120493

Ofetotse, E. L., Essah, E. A., & Yao, R. (2021). Evaluating the determinants of household electricity consumption using cluster analysis. *Journal of Building Engineering*, *43*, 102487. https://doi.org/https://doi.org/10.1016/j.jobe.2021.102487

Quilumba, F. L., Lee, W.-J., Huang, H., Wang, D. Y., & Szabados, R. L. (2015). Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities. *IEEE Transactions on Smart Grid*, *6*(2), 911-918.

Rajabi, A., Eskandari, M., Jabbari Ghadi, M., Ghavidel, S., Li, L., Zhang, J., & Siano, P. (2019). A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications. *Energy and Buildings*, *203*, 109455. https://doi.org/https://doi.org/10.1016/j.enbuild.2019.109455

Wang, Y., Chen, Q., Kang, C., & Xia, Q. (2016). Clustering of electricity consumption behavior dynamics toward big data applications. *IEEE Transactions on Smart Grid*, *7*(5), 2437-2447.

Wen, L., Zhou, K., & Yang, S. (2019). A shape-based clustering method for pattern recognition of residential electricity consumption. *Journal of Cleaner Production*, *212*, 475-488. https://doi.org/https://doi.org/10.1016/j.jclepro.2018.12.067

Westermann, P., Deb, C., Schlueter, A., & Evins, R. (2020). Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data. *Applied Energy*, *264*, 114715. https://doi.org/https://doi.org/10.1016/j.apenergy.2020.114715

Appendix

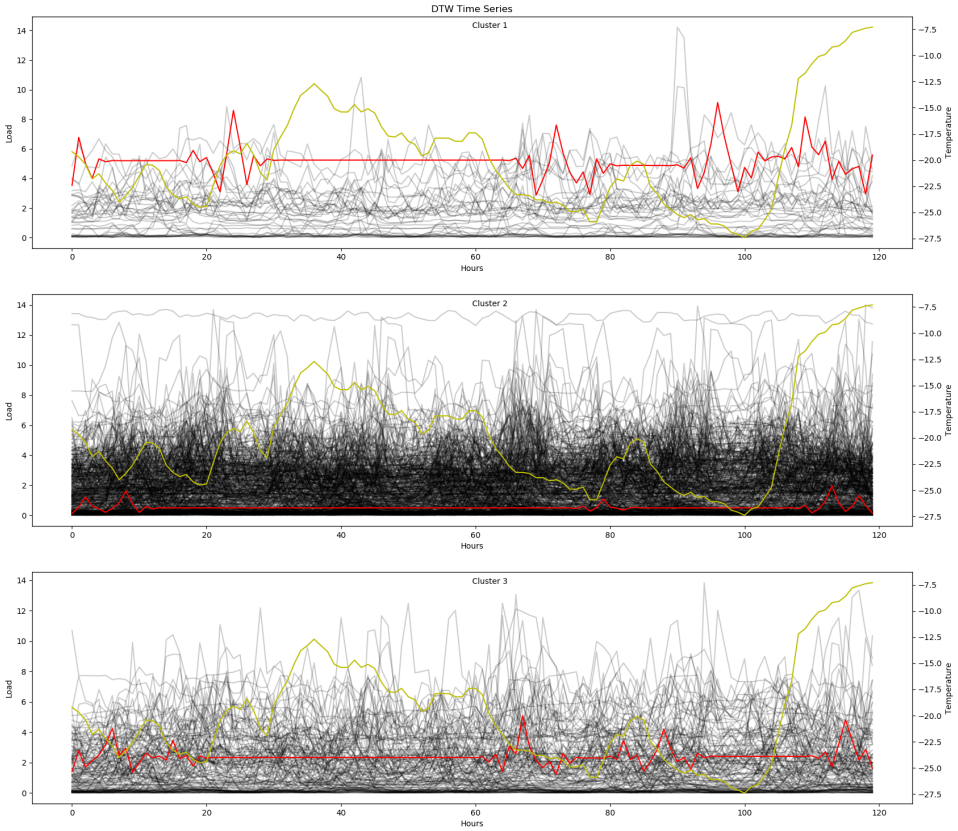

Figure 16. 3 clusters obtained from the original dataset.

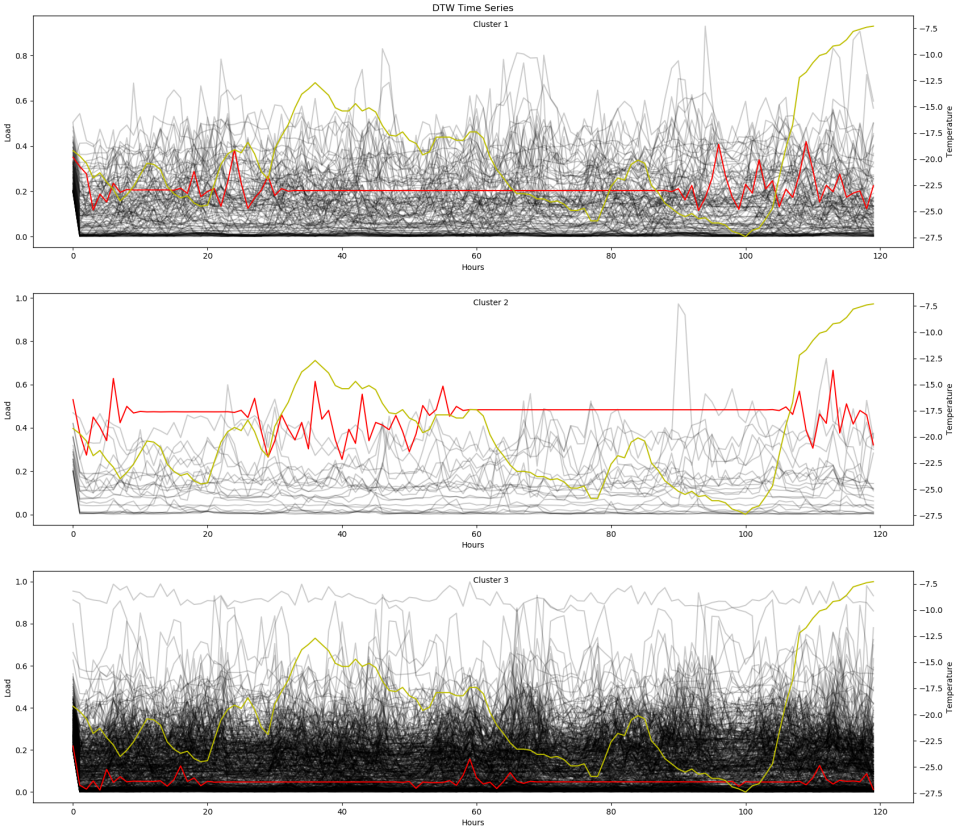

Figure 17. 3 clusters obtained from the scaled dataset.

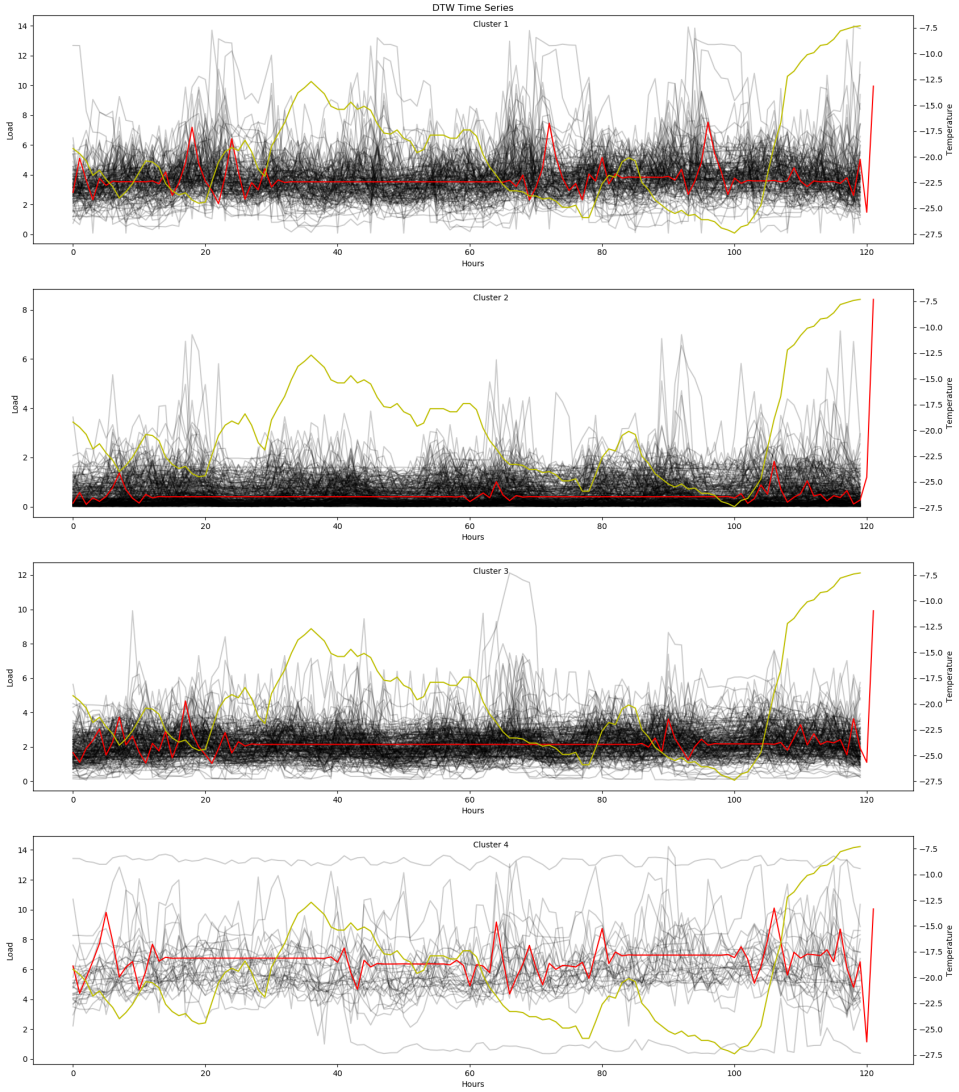

Figure 18. 4 clusters obtained from the original dataset.

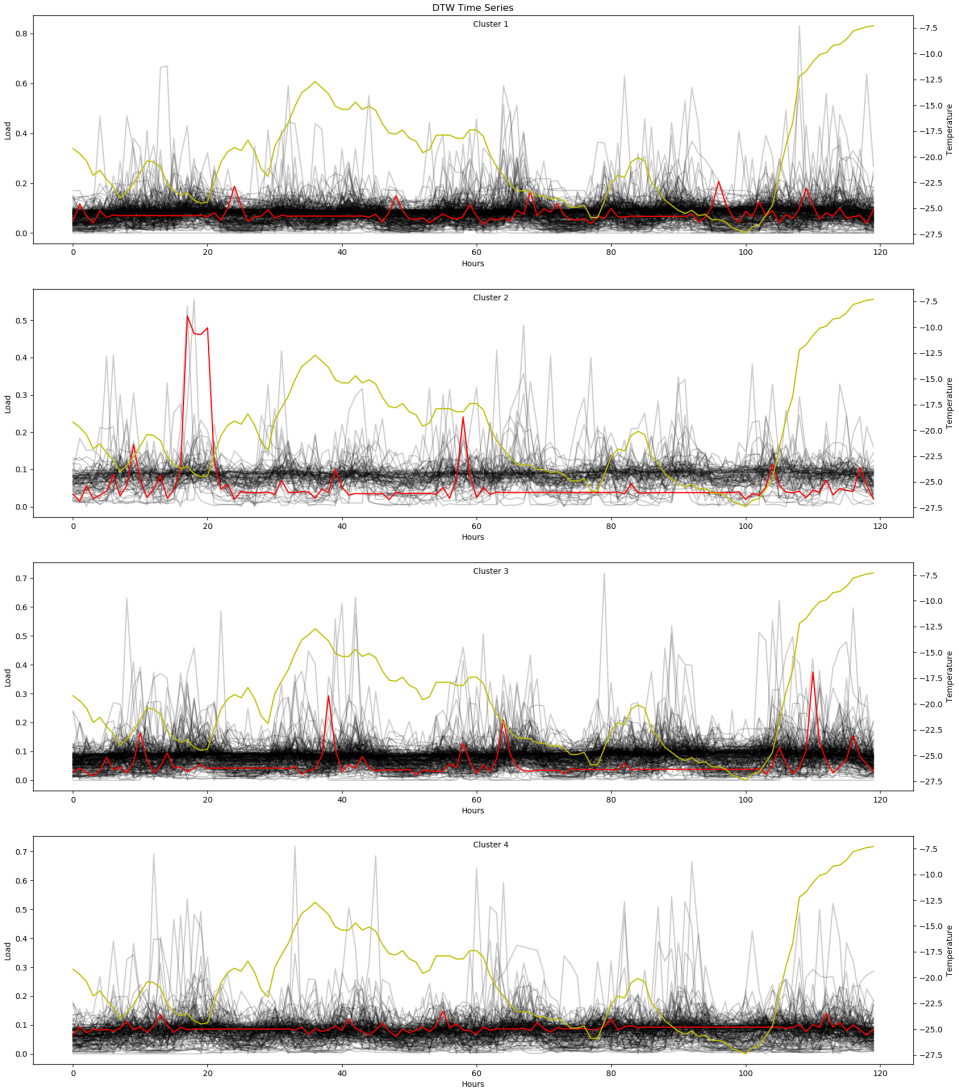

Figure 19. 4 clusters obtained from the normalized dataset.

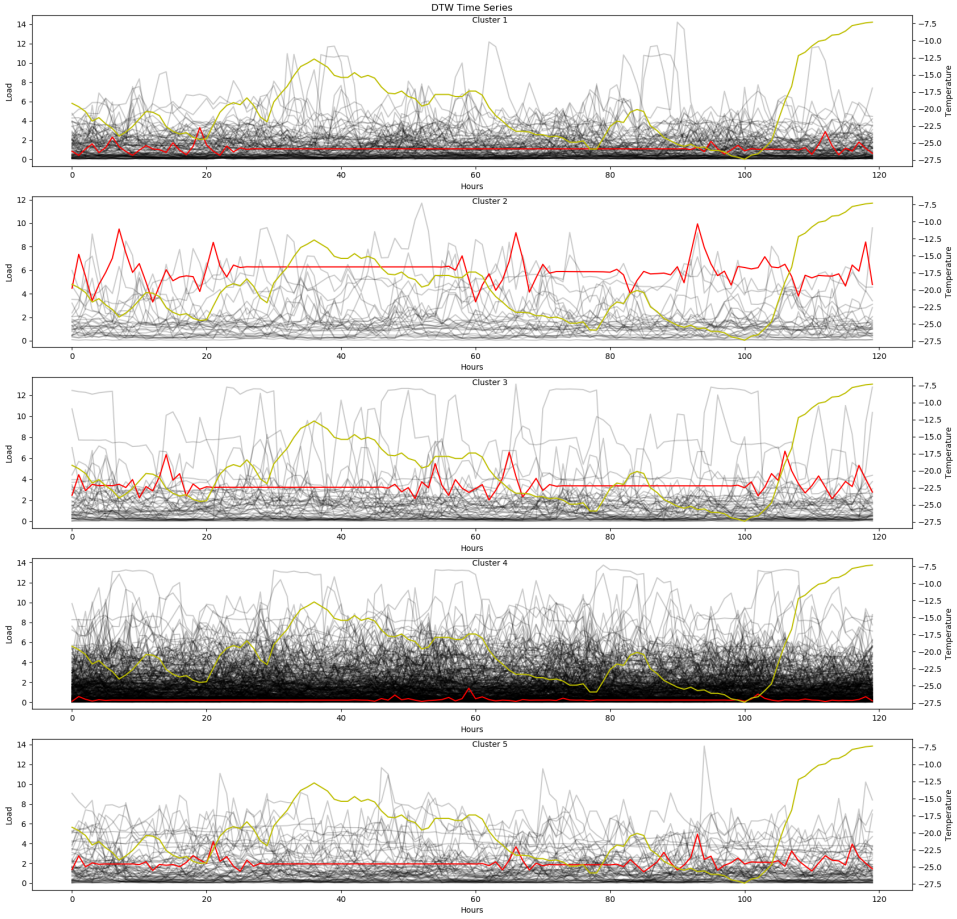

Figure 20. 5 clusters obtained from the original dataset.

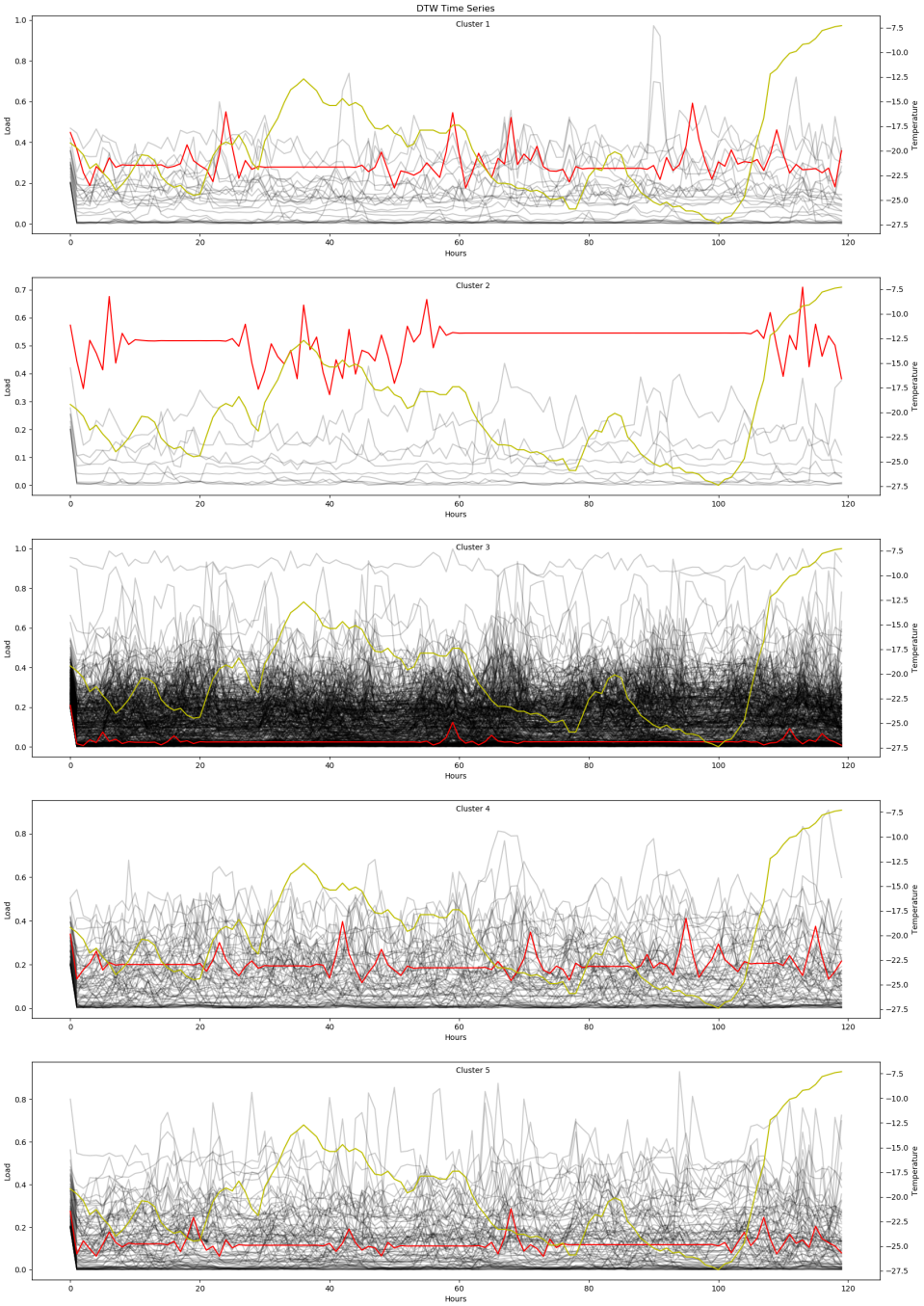

Figure 21. 5 clusters obtained from the scaled dataset.

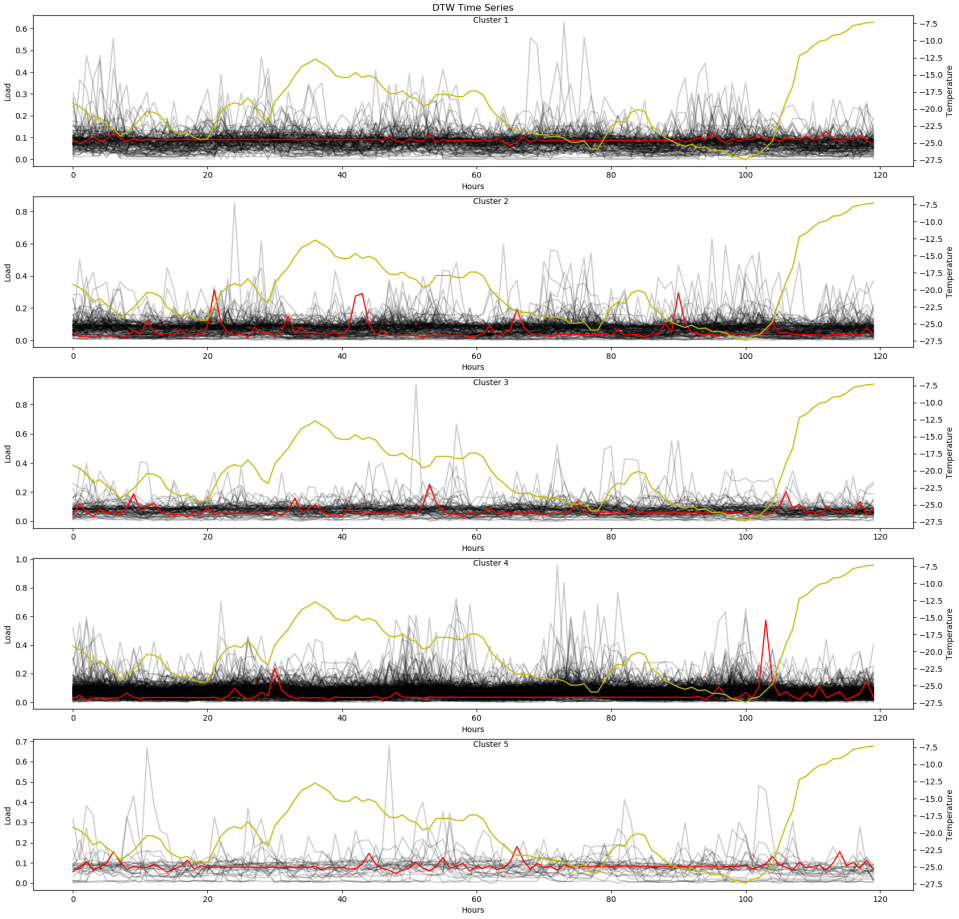

Figure 22. 5 clusters obtained from the normalized dataset.