POLITECNICO
DI TORINO

enel endesa

Characterising Electricity Consumption Over Time for Residential Consumers through cluster analysis

Tania Cerquitelli, Gianfranco Chicco, Evelina Di Corso, Francesco Ventura,

Giuseppe Montesano, Anita Del Pizzo, Mirko Armieto

Alicia Mateo González, Eduardo Martin Sobrino, Andrea Veiga Santiago

Introduction

- The knowledge of the ***consumers' electrical behaviour*** is a key aspect for various electrical system operators
  - distribution system operators
  - aggregators
  - retailers
- Research challenge
  - discovering consistent ***groups of consumers***
  - with ***common patterns of electricity consumption*** in ***a given time period***

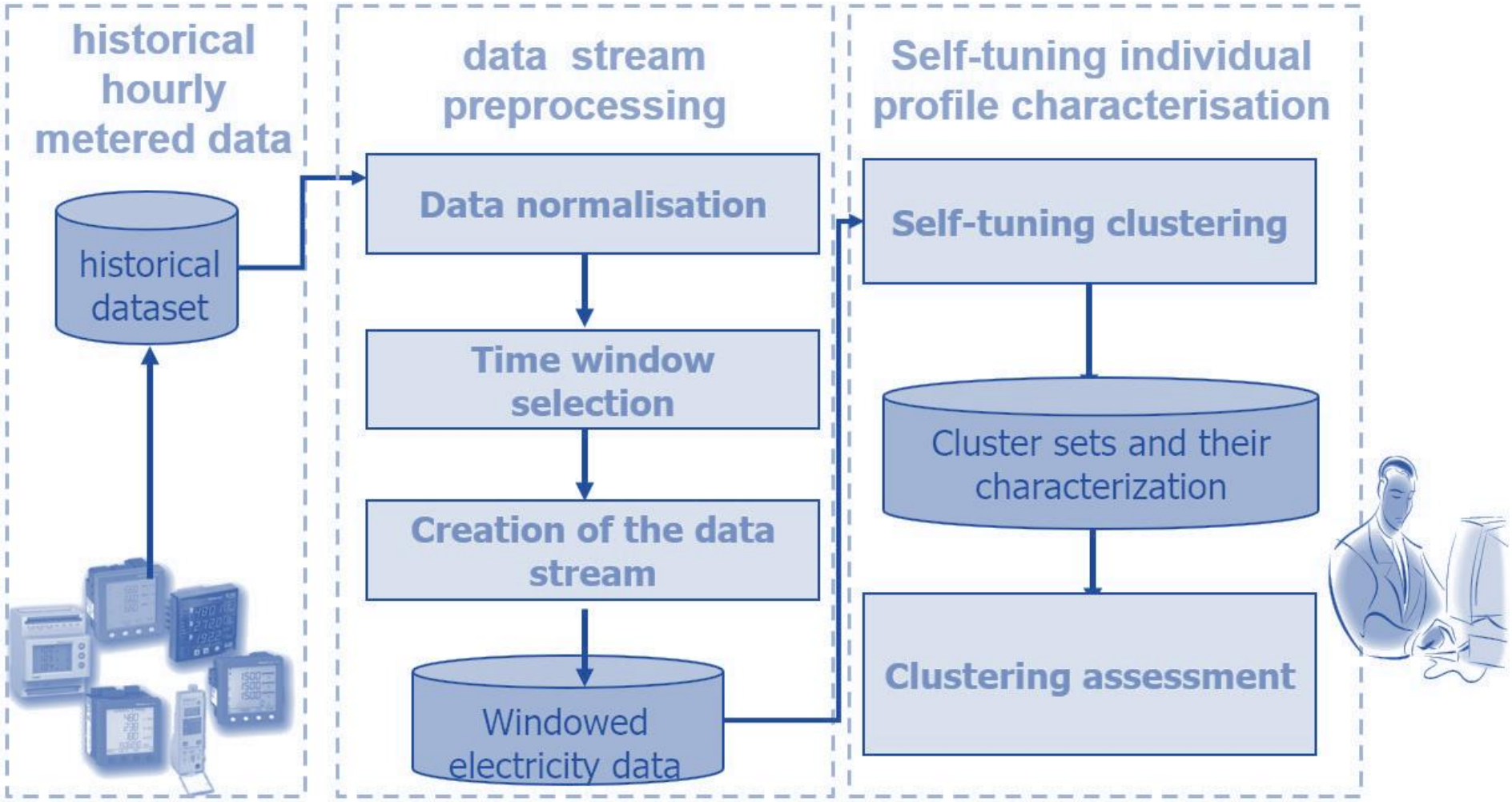
POLITECNICO DI TORINO

CONDUCTS (CONsumption DUration Curve Time Series)

- exploits ***data stream processing***
- implements ***unsupervised machine learning*** to identify
  - *patterns of individual electricity consumption*
  - *consumer behaviour over time*
- provides ***time windows analysis*** for
  - Weekdays
  - Weekend and special days
- exploits ***state-of-the-art distributed computing frameworks***
  - It quickly analyses *very large energy data collections*
  - It supports *parallel and scalable processing*

POLITECNICO DI TORINO

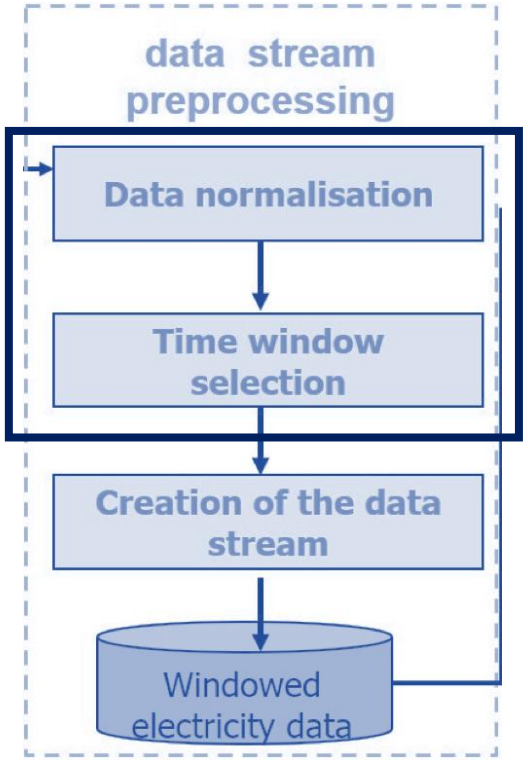
CONDUCTS (CONsumption DUration Curve Time Series)



POLITECNICO DI TORINO

Data stream pre-processing

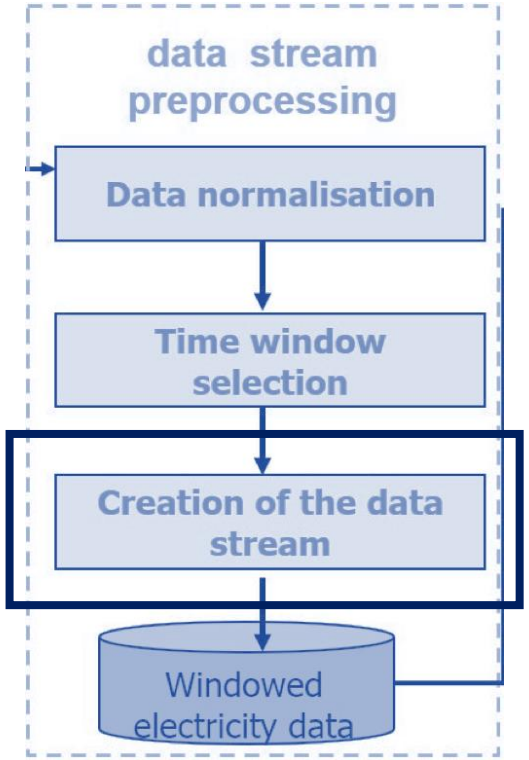
- Data normalisation
  - focused on ***discovering shape-based patterns***, removing the effect of the consumption level
  - normalisation ***with respect to the contract power***
    - it allows to easily identify outliers (consumption greater than one after the normalisation)
- Time window selection
  - time window length should be able to ***capture different external variables***
    - Temperature
    - weather conditions
  - *weekdays and non-common days* have been analysed *separately*



POLITECNICO DI TORINO

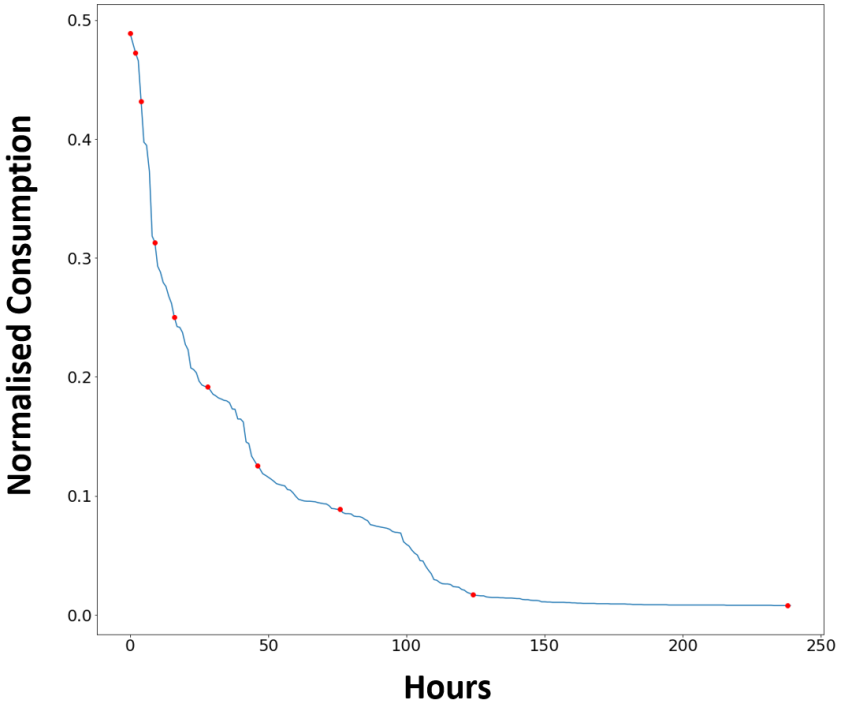
Data stream pre-processing

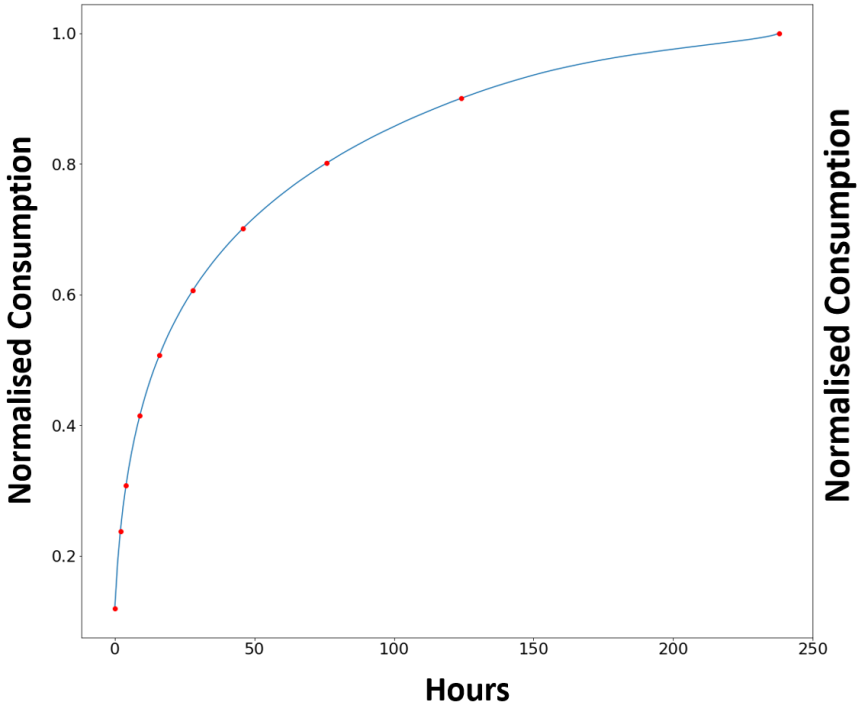
- Creation of the **duration curve of the consumption**
  - Ordering
    - The hourly data for each consumer are ordered in descending values
  - Variations computation
    - The average variations between each value and the following one have been calculated
  - Cumulative of the variations
    - the variations have been summed up sequentially
    - values ranging from 0 to 1 on the vertical axis
    - Interpreted as a cumulative distribution function (CDF)
  - Cut points selection
    - 9 points on the horizontal axis corresponding to the last 9 deciles have been identified as cut points
    - The first decile is excluded because of its low variations

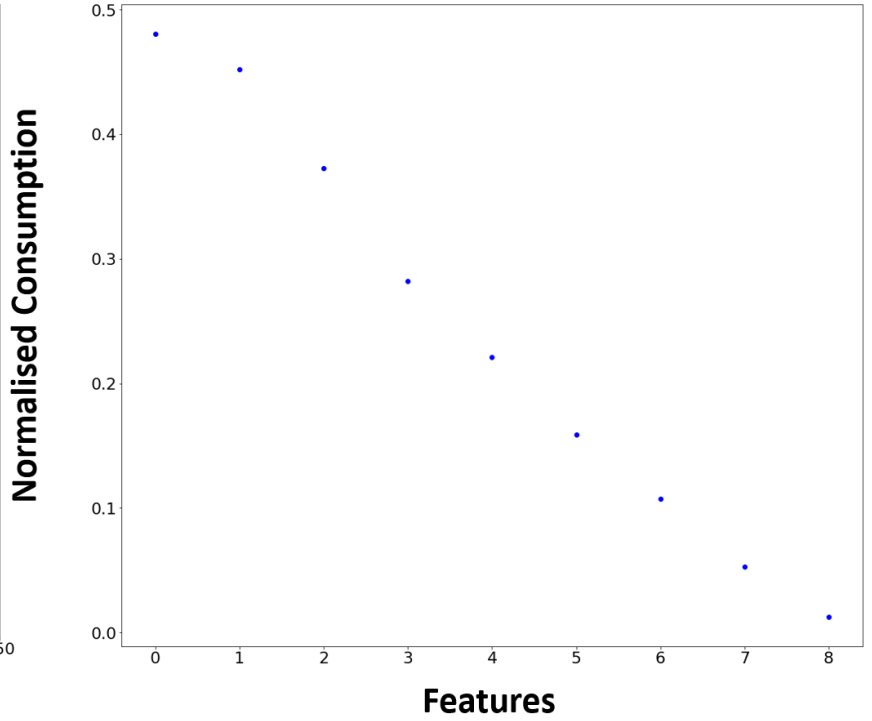


POLITECNICO DI TORINO

Creation of the consumption duration curve







Ordering

The hourly data for each consumer are ordered in descending values

Variations computation

The average variations between each value and the following one have been calculated

Cumulative of the variations

The variations have been summed up sequentially

Values ranging from 0 to 1 on the vertical axis

Interpreted as a cumulative distribution function (CDF)

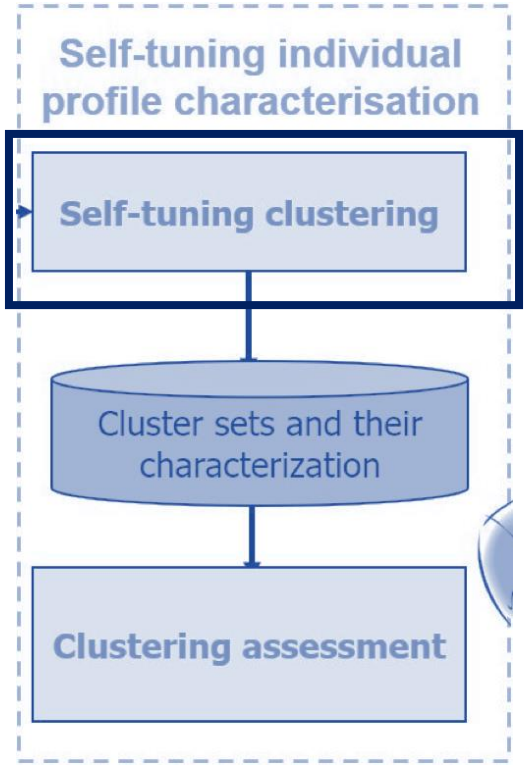
Cut points selection

9 points on the horizontal axis corresponding to the last 9 deciles have been identified as cut points

POLITECNICO DI TORINO

Self-tuning individual profile characterisation

- Self-tuning clustering algorithm includes
  - ***K-means algorithm*** with ***Euclidean distance***



POLITECNICO DI TORINO

K-means algorithm

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified

- ‘Closeness’ in CONDUCTS is measured by Euclidean distance
  - With time dependent data the Euclidean distance is not appropriate
  - Cut points of each duration curve are monotonically decreasing thus the selection of the Euclidean distance is reasonable

Evaluating the K-means partitions

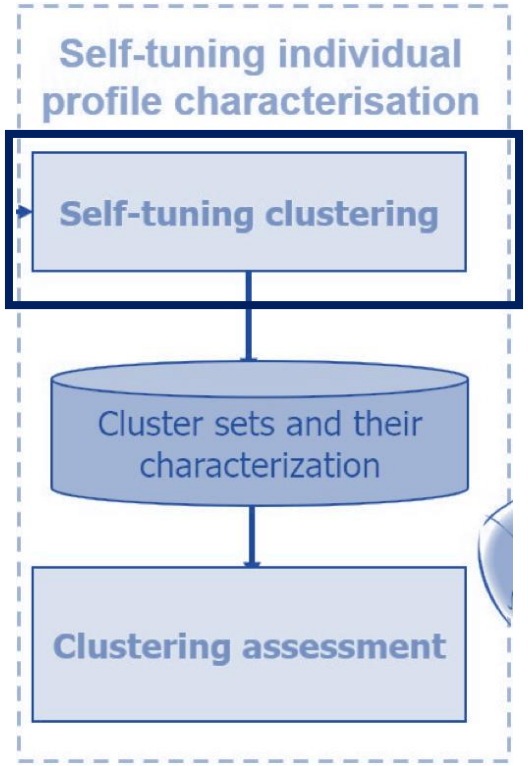
- Most common measure is Sum of Squared Error (SSE)
- For each point, the error is the distance to the nearest cluster
- To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- x is a data point in cluster Ci and mi is the representative point for cluster Ci
  - can show that mi corresponds to the center (mean) of the cluster
- Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase K, the number of clusters
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Self-tuning individual profile characterisation

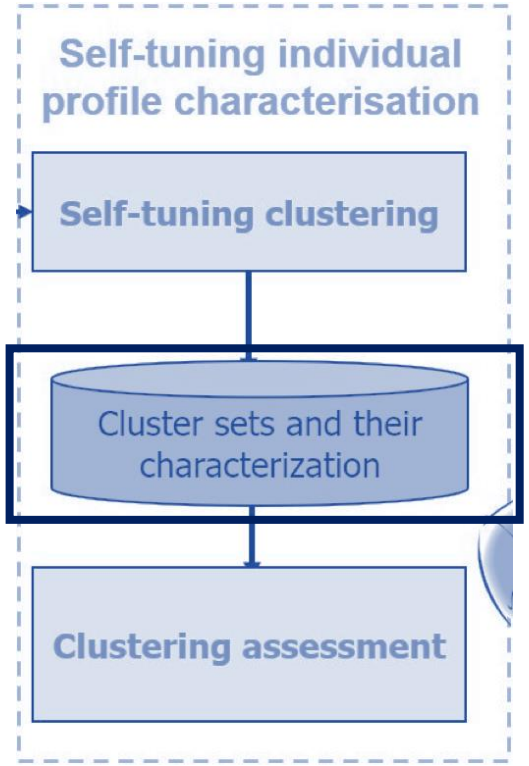
- Self-tuning clustering algorithm includes
  - ***K-means algorithm*** with ***Euclidean distance***
  - A self-tuning strategy to automatically discover the desired number of clusters
    - ***K is automatically configured for each time window***
      - **Elbow of SSE against K**
        - Plotting the quality measure trend (e.g., SSE) against K
        - Choosing the value of K
          - the gain from adding a centroid is negligible
          - The reduction of the quality measure is not interesting anymore
          - SSE reduction is worth enough compared with the number of clusters



POLITECNICO DI TORINO

Self-tuning individual profile characterisation

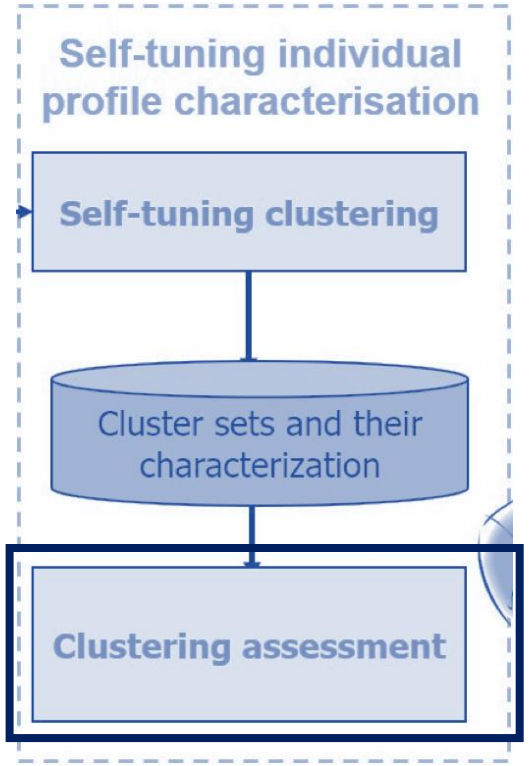
- Clustering characterisation
  - ***Centroids-based characteristics***
    - graphically show an overview of the discovered cluster sets
    - plot of the duration curves corresponding to each cluster centroids
  - ***Scatter plot of the duration curves***
    - relations between the maximum hourly consumption and the average consumption in each cluster
  - ***Boxplot distribution***
    - for each cut point
    - for averaged daily time series



POLITECNICO DI TORINO

Self-tuning individual profile characterisation

- Clustering assessment
  - It evaluates the ***ability of the CONDUCTS*** engine to correctly identify ***groups of individual consumers***
  - The ***silhouette index*** evaluates the quality of the cluster models
    - It measures **intra-cluster cohesion** and **inter-cluster separation**
    - It takes values in [-1,1]
      - Positive values -> good partitioning
      - Negative values -> bad partitioning
    - It could be computed for each consumer, for each cluster, and for the cluster set



POLITECNICO DI TORINO

The Silhouette index

- The silhouette index is a quality measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).
- The silhouette ranges from -1 to +1
  - a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring cluster.
- For each load pattern *i*, the silhouette is defined as:

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

where a_i is the average distance between *i* and the other load patterns in the same cluster, while b_i is the lowest average distance between the load pattern *i* and each one of the other clusters (not containing the load pattern *i*)

Experimental results: Addressed issues

- ***Self-tuning strategy to automatically set K***
  - SSE trend analysis
- ***Clustering assessment***
  - Silhouette values for clusters in each time window
- ***Clustering characterization***
  - Number of consumers for each cluster
  - Centroid distribution for each cluster
  - Cut points boxplot distribution versus daily consumption boxplot distribution
  - Relations between the maximum and the average normalised hourly energy consumption for each cluster
- ***Sensitivity analysis and comparison wrt state-of-the-art approaches***

POLITECNICO DI TORINO

Experimental setting

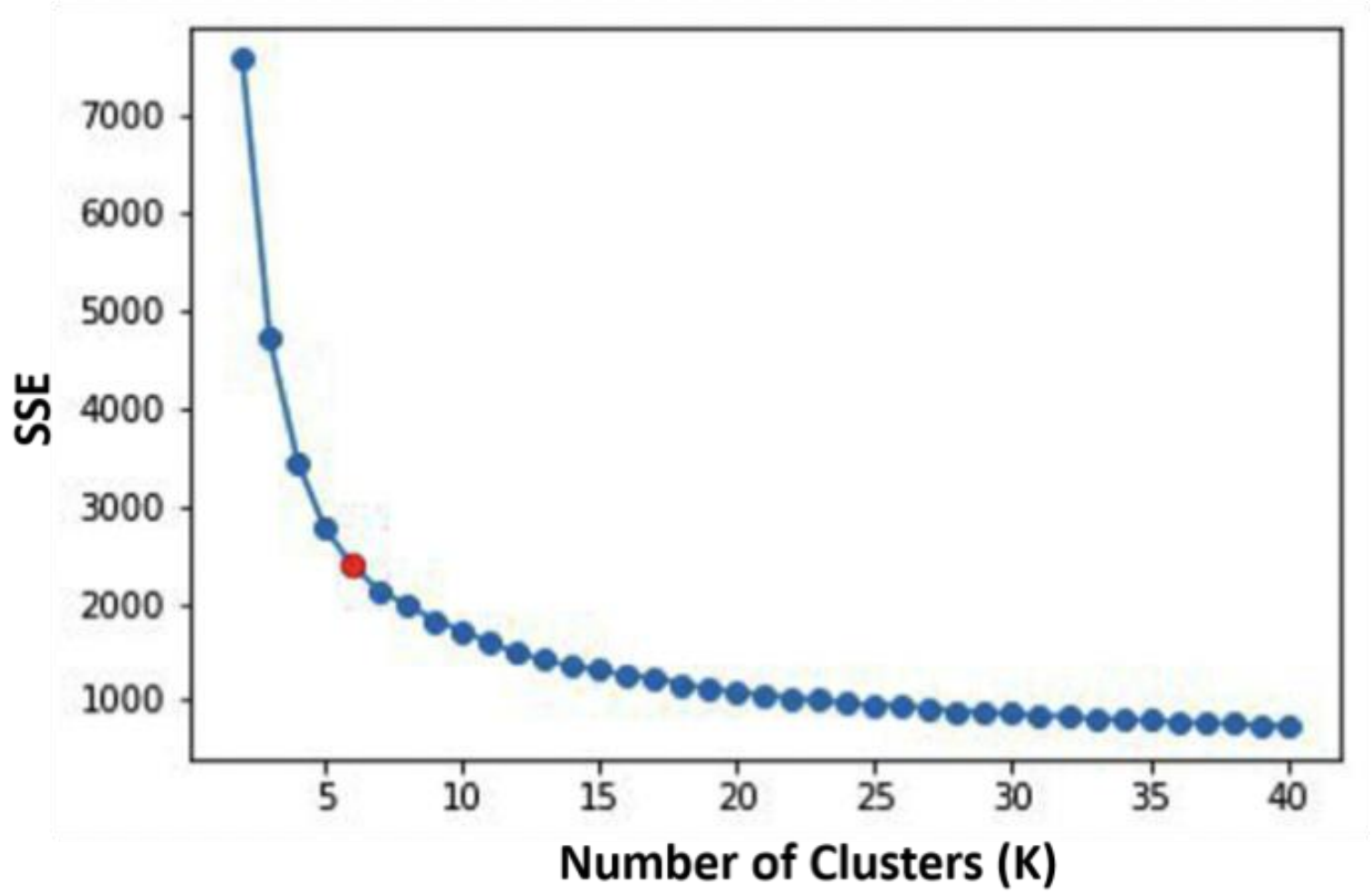
- Real dataset
  - ***real hourly-metered data***
    - related to ***565,662 Spanish residential consumers***
    - collected during one year (from 2016-05-01 to 2017-04- 30)
  - Availability of individual contract power
    - used to normalise the data
- CONDUCTS project
  - Developed in Scala
  - Spark framework
    - Engine for parallel and distributed big-data analytics
  - Spark MLlib
    - Scalable machine learning library for Spark

Experimental setting

- ***Different time windows*** in order to ***separate working and non-working days***
  - Two weeks time window for the working days
  - Monthly time window for non-working days and special days
- 38 time windows
  - ***26*** defined by ***working days***
  - ***12*** defined by ***non-working days***
- Results for the time window related to the ***first two weeks of June*** are discussed

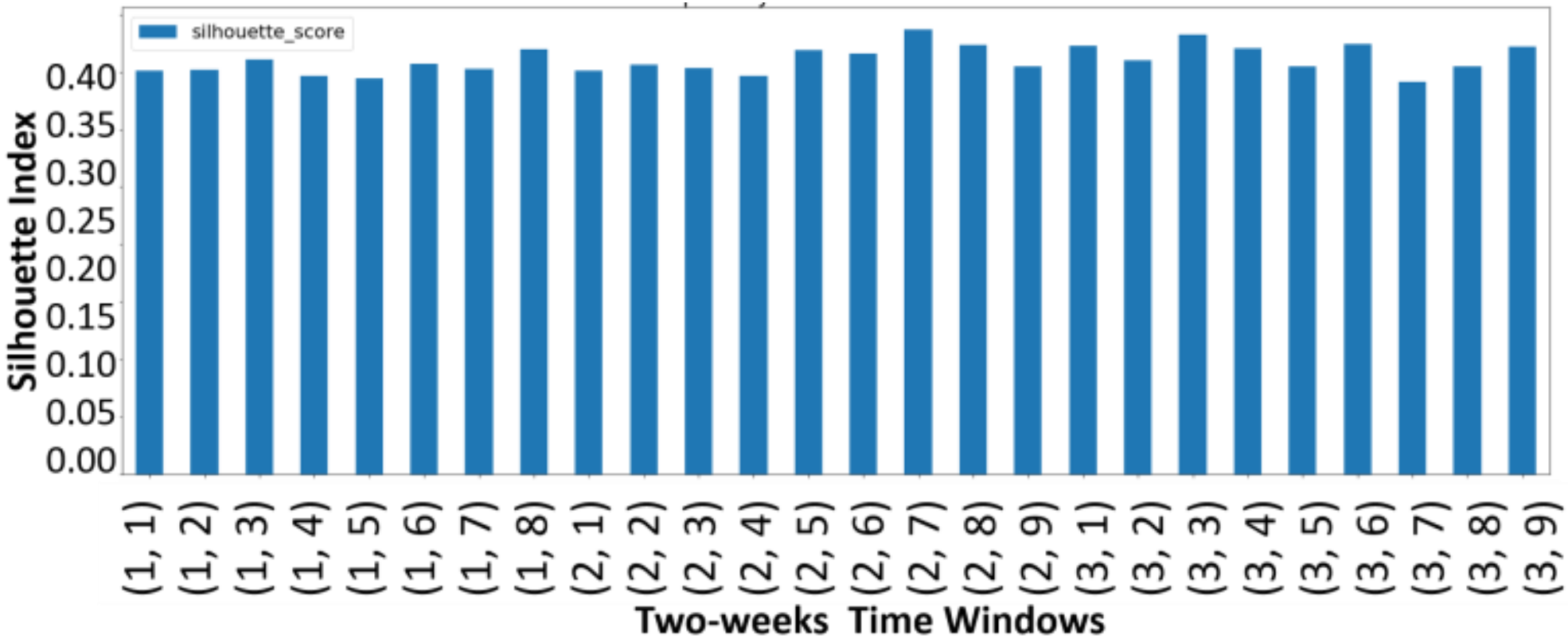
POLITECNICO DI TORINO

SSE trend analysis to automatically set K



POLITECNICO DI TORINO

Silhouette values for clusters in each working day time window

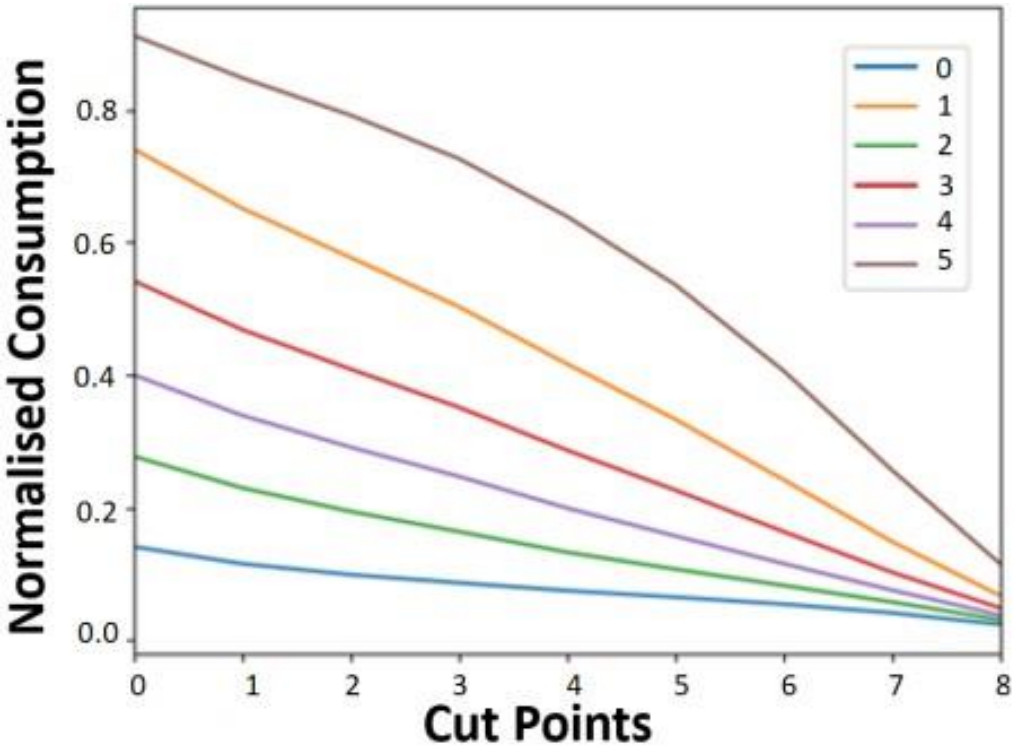


POLITECNICO DI TORINO

Cluster set characterisation

Cluster ID	*Size*
0	108,993
1	36,489
2	174,270
3	90,207
4	147,856
5	7,847
Total number of consumers	***565,662***

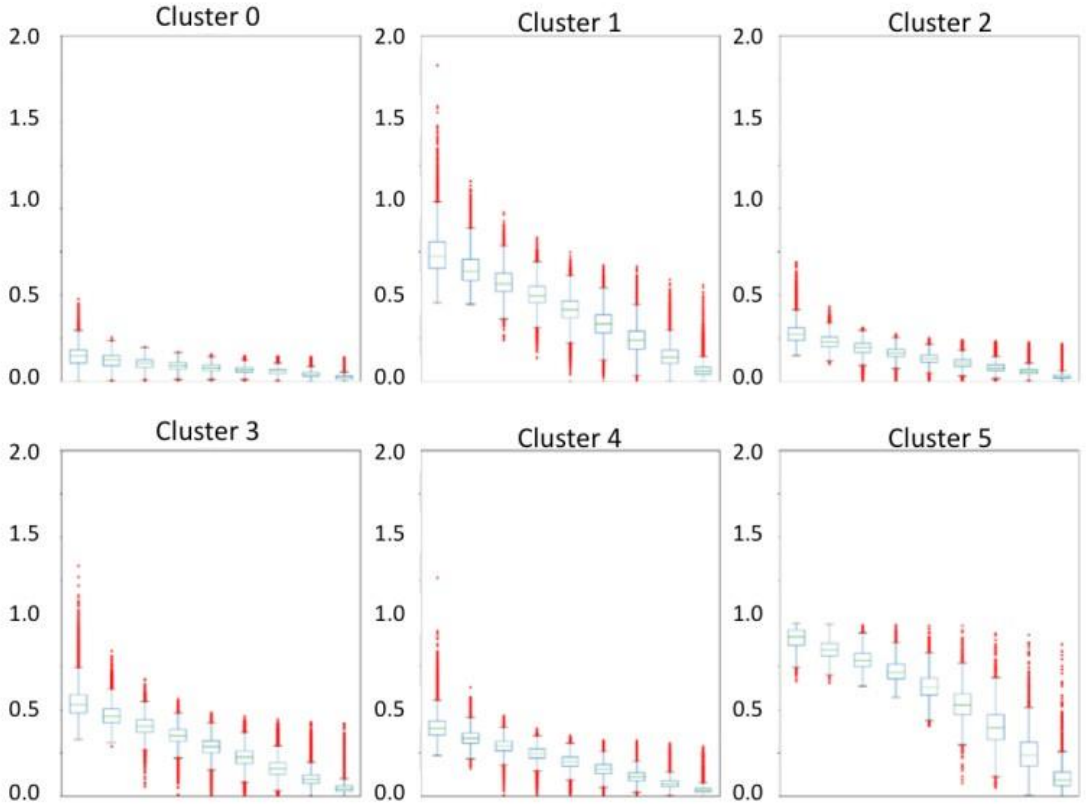
Centroid distribution (*energy utilisation factor - the energy consumption to contract power ratio*) for each cluster



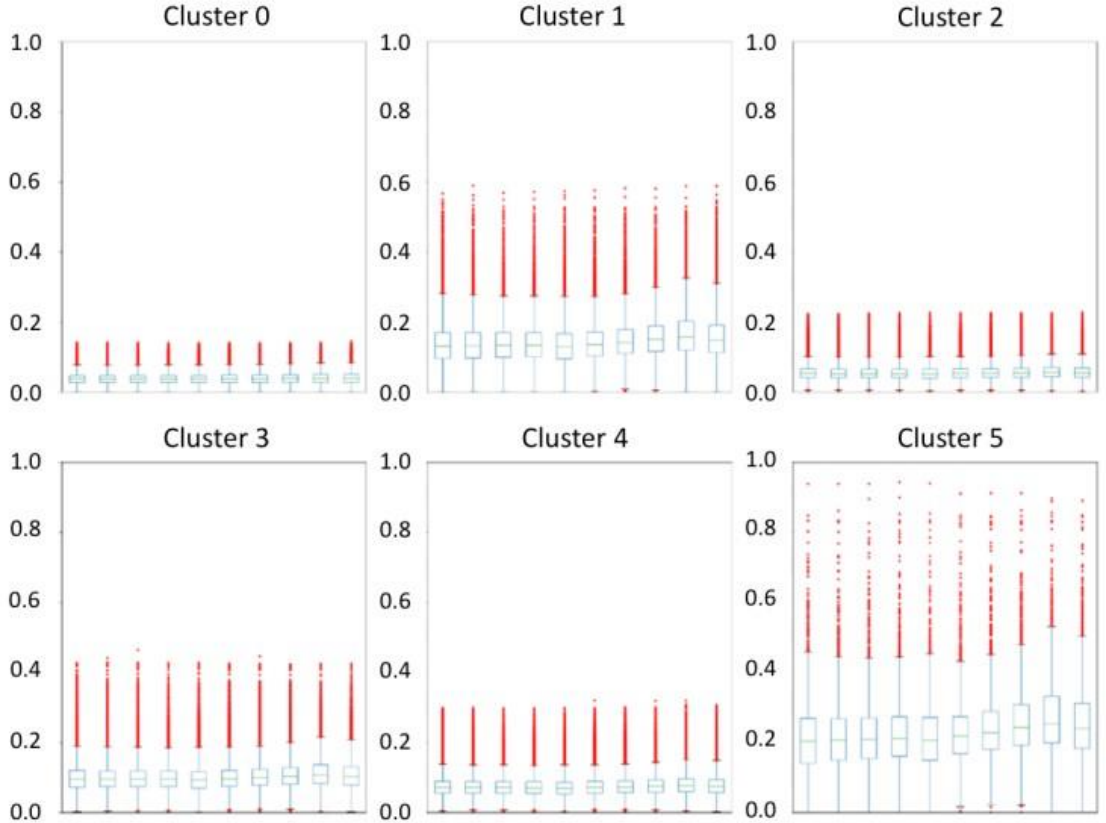
POLITECNICO DI TORINO

Cluster set characterisation

Cut points boxplot distribution

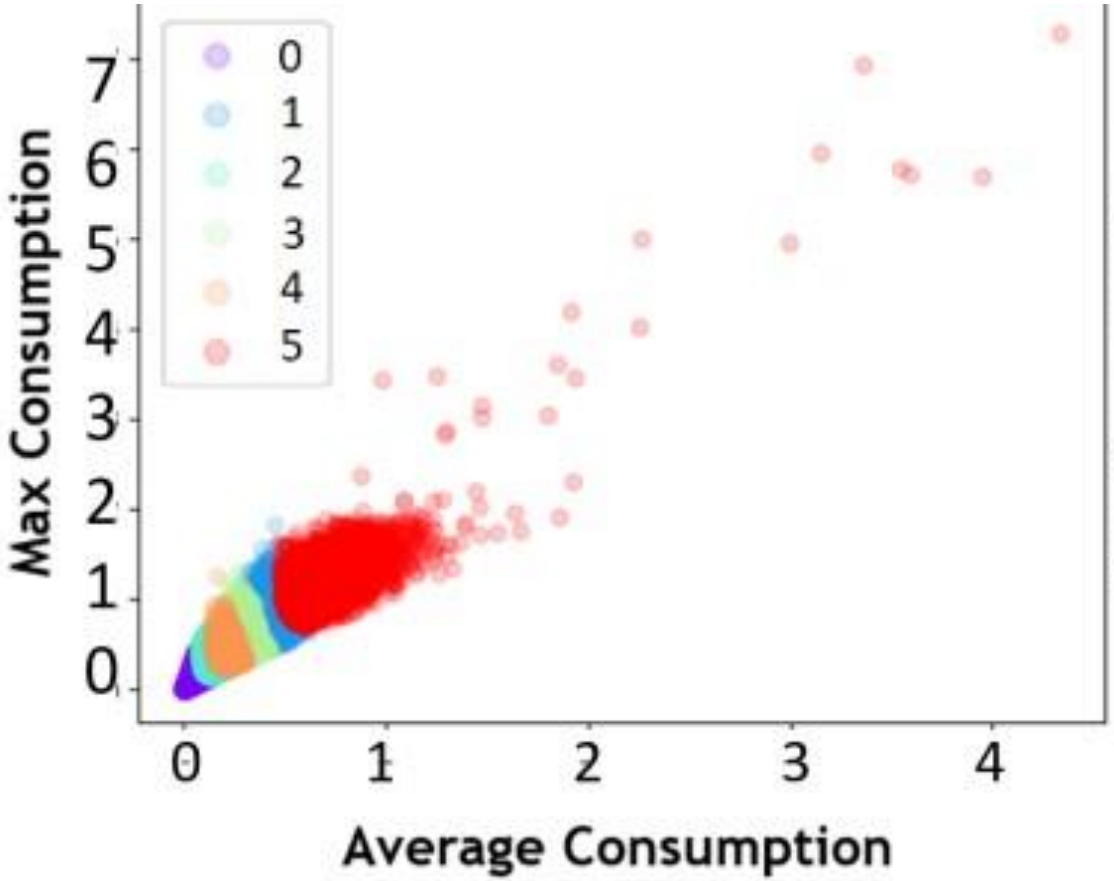


Daily consumption boxplot distribution

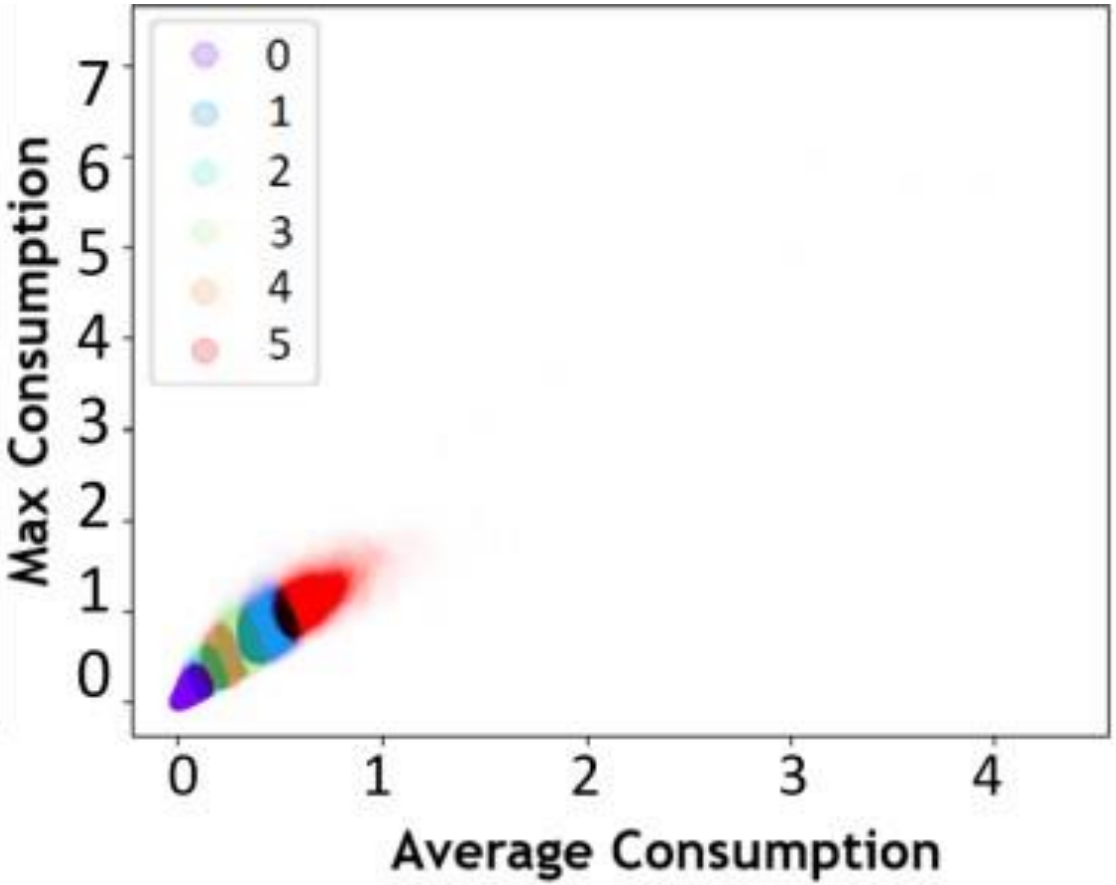


POLITECNICO DI TORINO

Scatter plot for the average and maximum normalised hourly energy consumption of each consumer



colour intensity a = 0.2



colour intensity a = 0.002

POLITECNICO DI TORINO

Further experiments

- We focused on a sampled dataset including 10,000 consumers
  - to configure the CONDUCTS methodology
    - Data normalisation
    - Distance measure
    - Data input of the cluster analysis
  - to compare CONDUCTS with respect to the state-of-the-art approach exploiting time series to drive the cluster analysis

- Silhouette-based indices are exploited to compare the quality of different partitions

Silhouette-based indices

- Two indicators are based on the previous definition:
  - The **average silhouette index** (ASI)

$$ASI = \frac{1}{N}\sum_{k=1}^{K}\sum_{i \in L_k} s_i \qquad s_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

  - The **global silhouette index** (GSI)

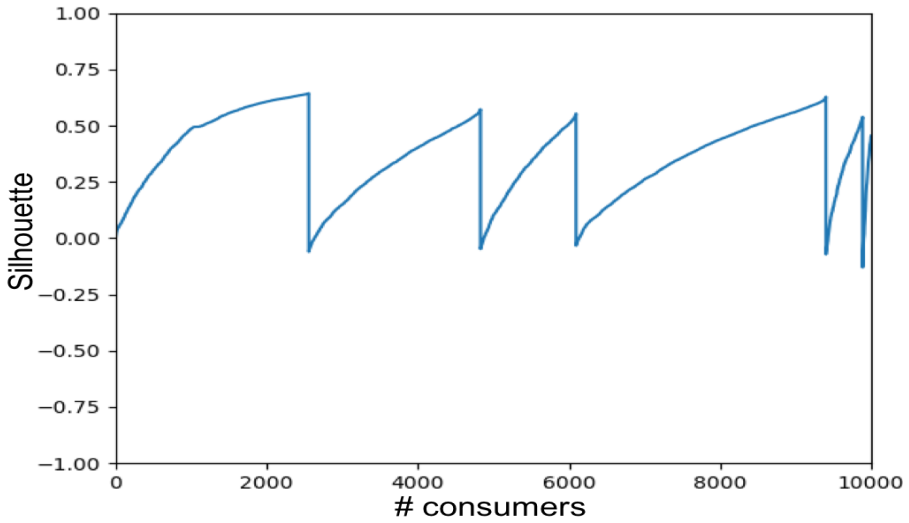
$$GSI(K) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|L_k|}\sum_{i \in L_k} s_i$$

Where L_k is the set of the patterns belonging to cluster $k = 1,...,K$; $|L_k|$ is the cardinality of cluster L_k (load patterns i belonging to the cluster L_k), and N is the total number of load patterns clustered (i.e., the number of consumers).

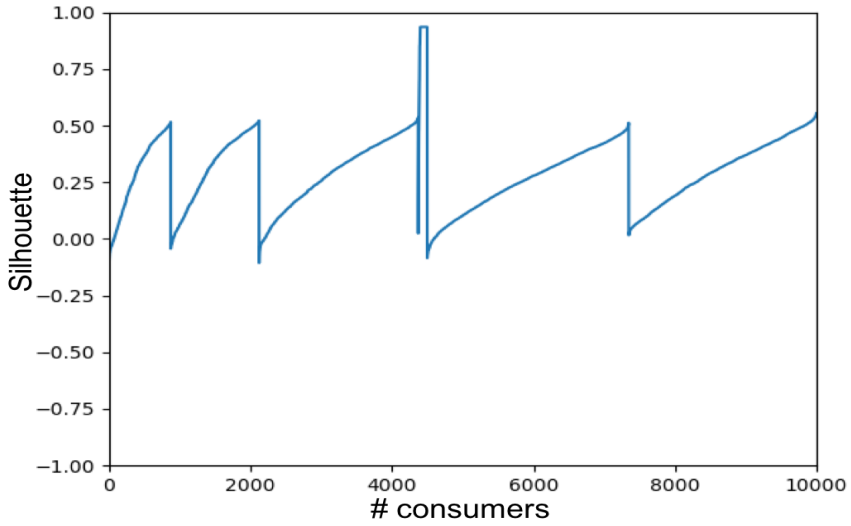
Id	Dataset	Normalization	K	Distance	ASI	GSI
C1	cut points_9	contract	6	dtw_d2	0.346	0.304
C2	cut points_9	contract	7	dtw_d2	0.337	0.245
C3	cut points_9	contract	6	dtw_d3	0.335	0.268
C4	cut points_9	contract	7	dtw_d3	0.325	0.255
C5	cut points_9	contract	6	Euclidean	**0.367**	**0.327**
C6	cut points_9	contract	7	Euclidean	0.343	0.307
C7	cut points_9	max-min	6	Euclidean	0.294	0.385
C8	cut points_9	max-min	7	Euclidean	0.273	0.352
H1	hourly_240	contract	6	Euclidean	-0.015	-0.016
H2	hourly_240	contract	7	Euclidean	**0.0421**	**0.060**
H3	hourly_240	contract	6	dtw_d10	-0.082	-0.089
H4	hourly_240	contract	7	dtw_d10	-0.060	-0.063

Configurations used for the K-means clustering and resulting Silhouette indicators ASI and GSI.

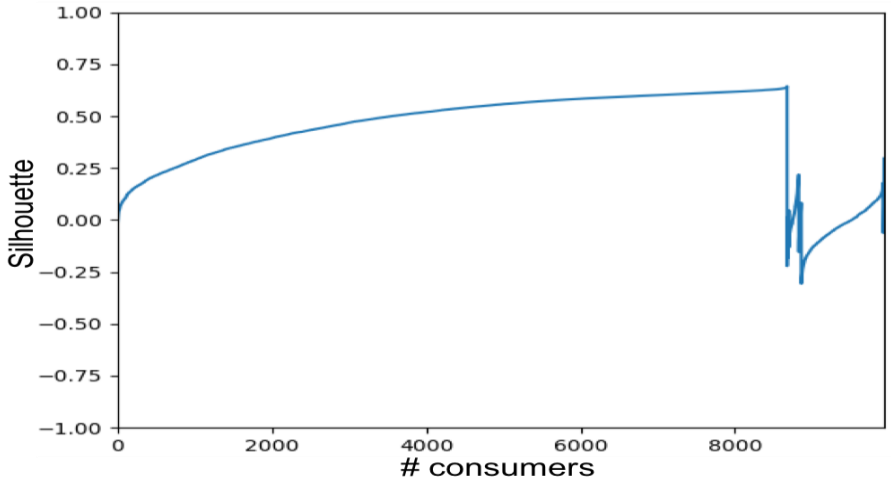
Silhouette plot for each consumer



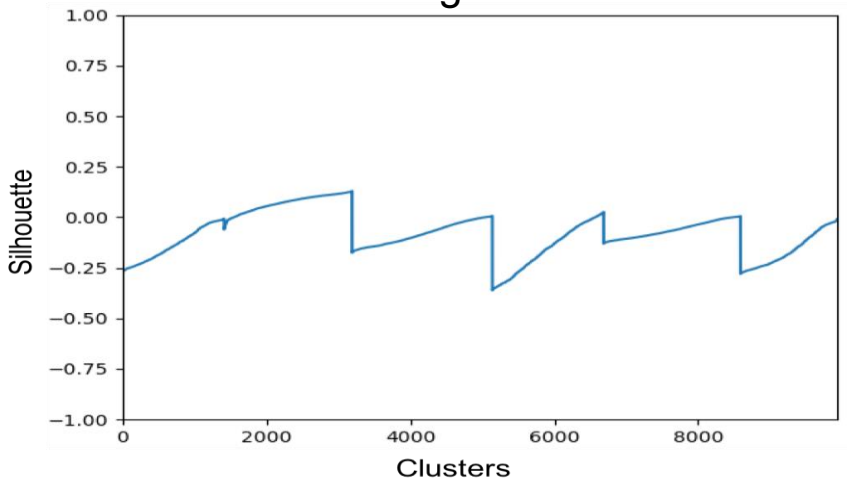
Configuration C5



Configuration C7



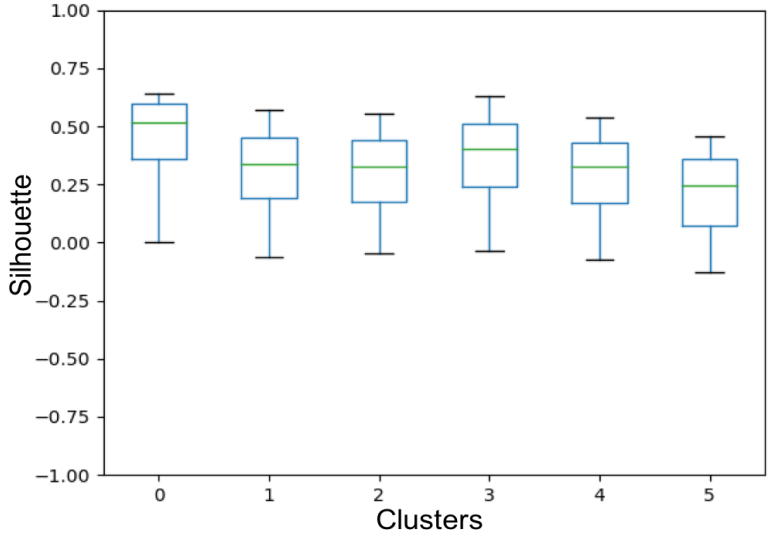
Configuration H2



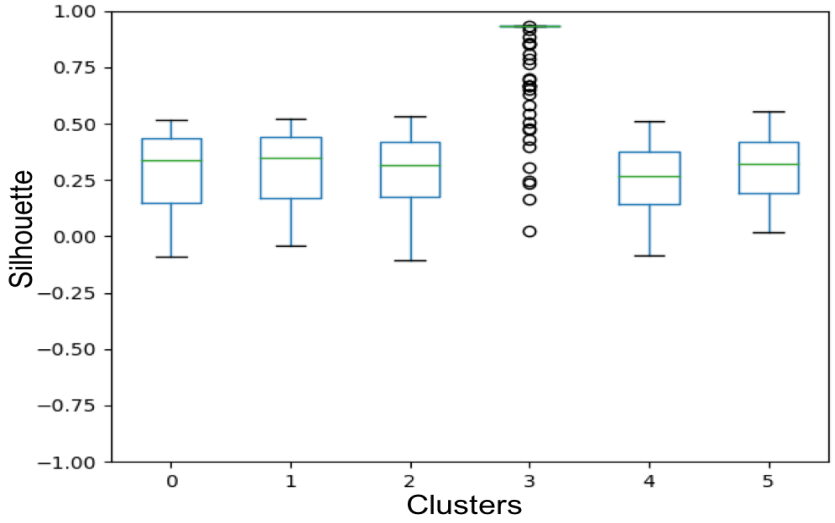
Configuration H3

POLITECNICO DI TORINO

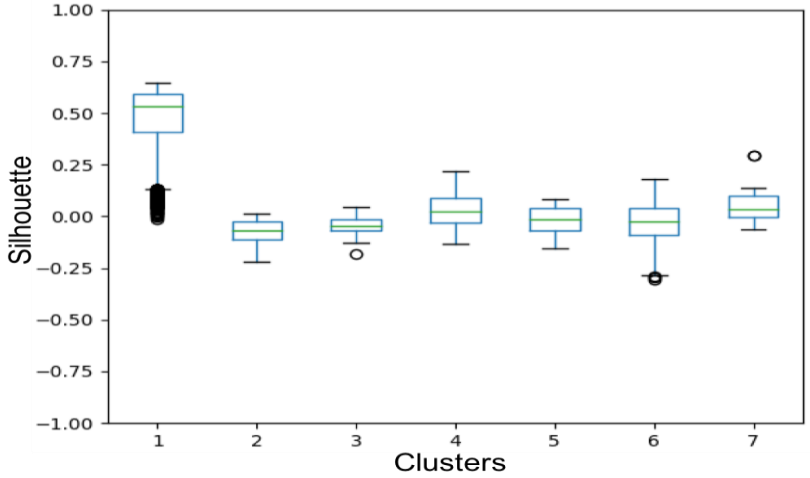
Silhouette boxplot for each cluster



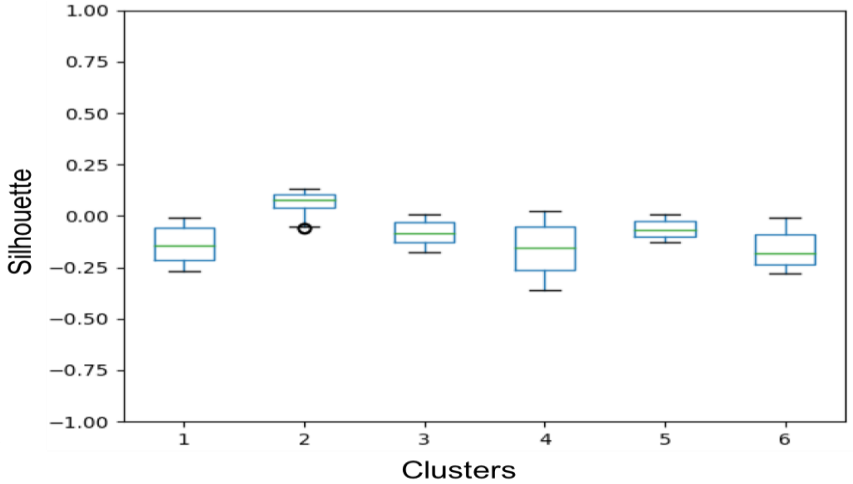
Configuration C5



Configuration C7



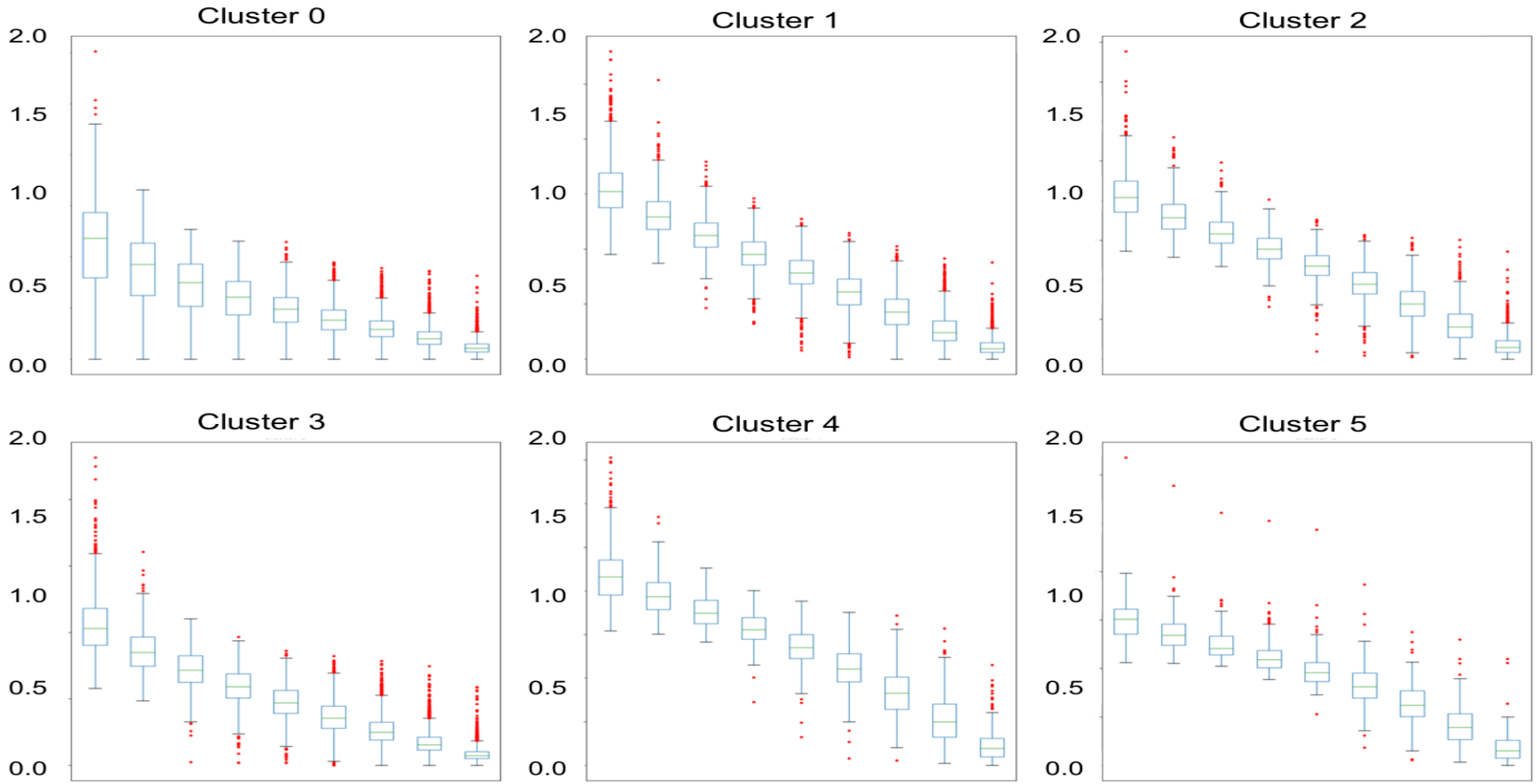
Configuration H2



Configuration H3

POLITECNICO DI TORINO

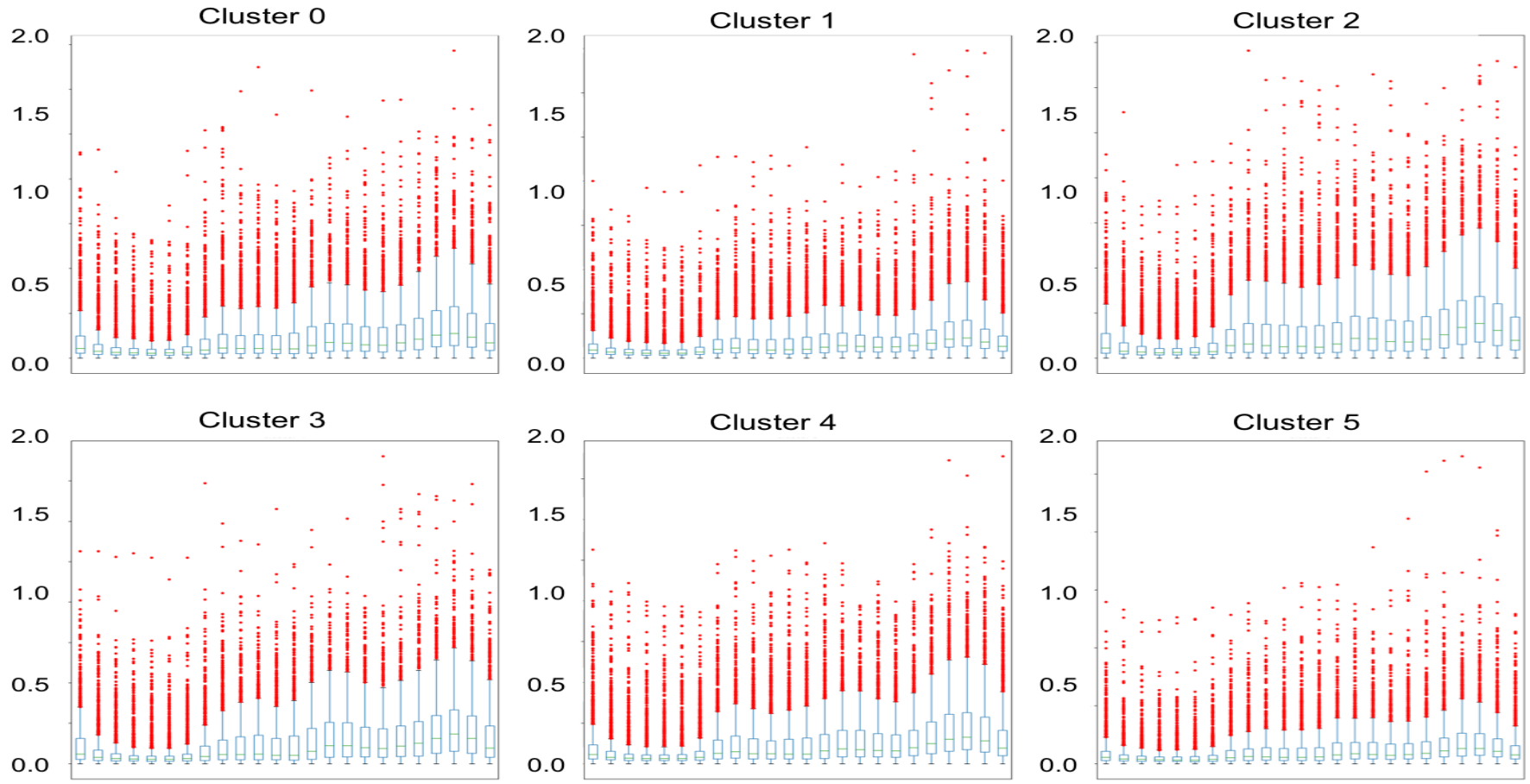
Configuration C5



Boxplot distribution of the nine cut points consumption for each cluster

POLITECNICO DI TORINO

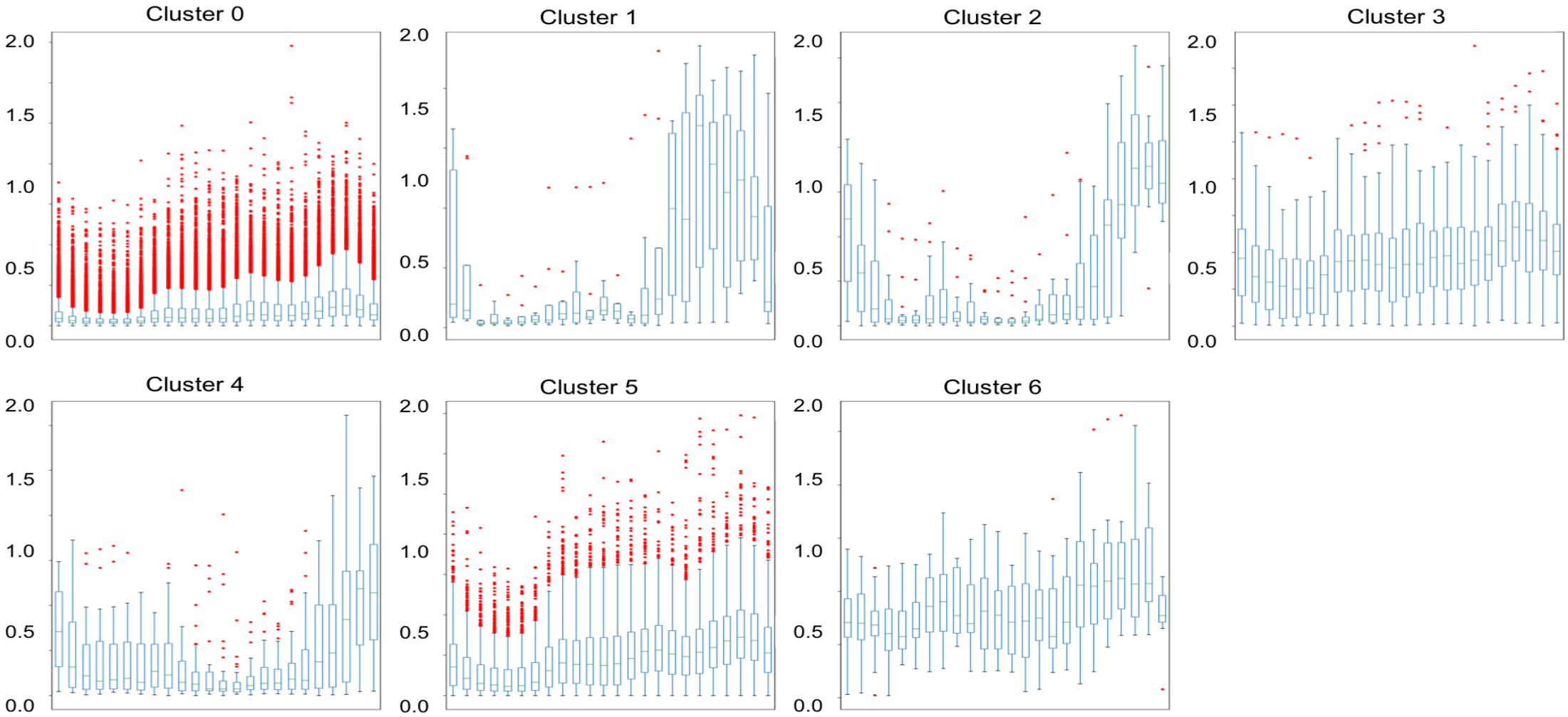
Configuration H2



Boxplot distribution of 24 hours of consumption (over 240 of the complete dataset) for each cluster

POLITECNICO DI TORINO

Configuration H3



Boxplot distribution of 24 hours of consumption (over 240 of the complete dataset) for each cluster

POLITECNICO DI TORINO

Conclusion and future works

- The proposed methodology is effective to correctly identify ***groups of individual consumers***
- Future works
  - build the ***electricity profile storyboard***
    - for each individual
    - using the clusters discovered on consecutive time windows
  - split the data into ***smaller slices*** to be ***analysed in further steps***

POLITECNICO DI TORINO

POLITECNICO
DI TORINO

enel endesa

Thank you!

Tania Cerquitelli, Gianfranco Chicco, Evelina Di Corso, Francesco Ventura,
Giuseppe Montesano, Anita Del Pizzo, Mirko Armieto
Alicia Mateo González, Eduardo Martin Sobrino, Andrea Veiga Santiago