Discovering frequent patterns in time series through unsupervised data mining techniques: the case of the energy profiling in buildings

Marco Savino Piscitelli

marco.piscitelli@polito.it

TEBE
TECHNOLOGY ENERGY BUILDING ENVIRONMENT

POLITECNICO DI TORINO

PyData

20/09/2018

Marco Savino Piscitelli - Education

Nov 2016 - ongoing

PhD - Energetics

Department of Energy – TEBE Research group - Politecnico di Torino

TOPIC:

Energy management in buildings through Data Analytics techniques

Scientific Tutors: Prof. Alfonso Capozzoli, Prof. Marco Perino

Feb 2015 - Sep 2016

Post-graduate grant researcher

Department of Energy – TEBE Research group - Politecnico di Torino

TOPIC:

- ***Sensitivity and advanced diagnostic analysis on buildings energy performance***
- ***Sensitivity and uncertainty analysis on health care facilities energy performance***

Scientific Tutor: Alfonso Capozzoli

Dec 2014

Master Degree in Building Systems Engineering

Politecnico di Torino

THESIS TITLE:

- ***Development of empirical models for estimating the energy demand of health care facilities and uncertainty analysis application***

Supervisor: Alfonso Capozzoli

Nov 2012

Bachelor Degree in Building Engineering

Politecnico di Bari

Technology Energy Building Environment Research Group (TEBE)

Technology
Energy
Building
Environment

TEBE RESEARCH TOPICS

- *Applied Acoustics*
- *Building energy performance*
- *Building envelope technologies*
- *HVAC technologies*
- *Indoor environmental quality*
- *Lighting*
- *Building Automation and Energy Data Analytics (BAEDA)*

Associate Professor
Alfonso Capozzoli

PhD student
Marco Savino Piscitelli
Topics:
DSS for building energy efficiency, energy classification and profiling

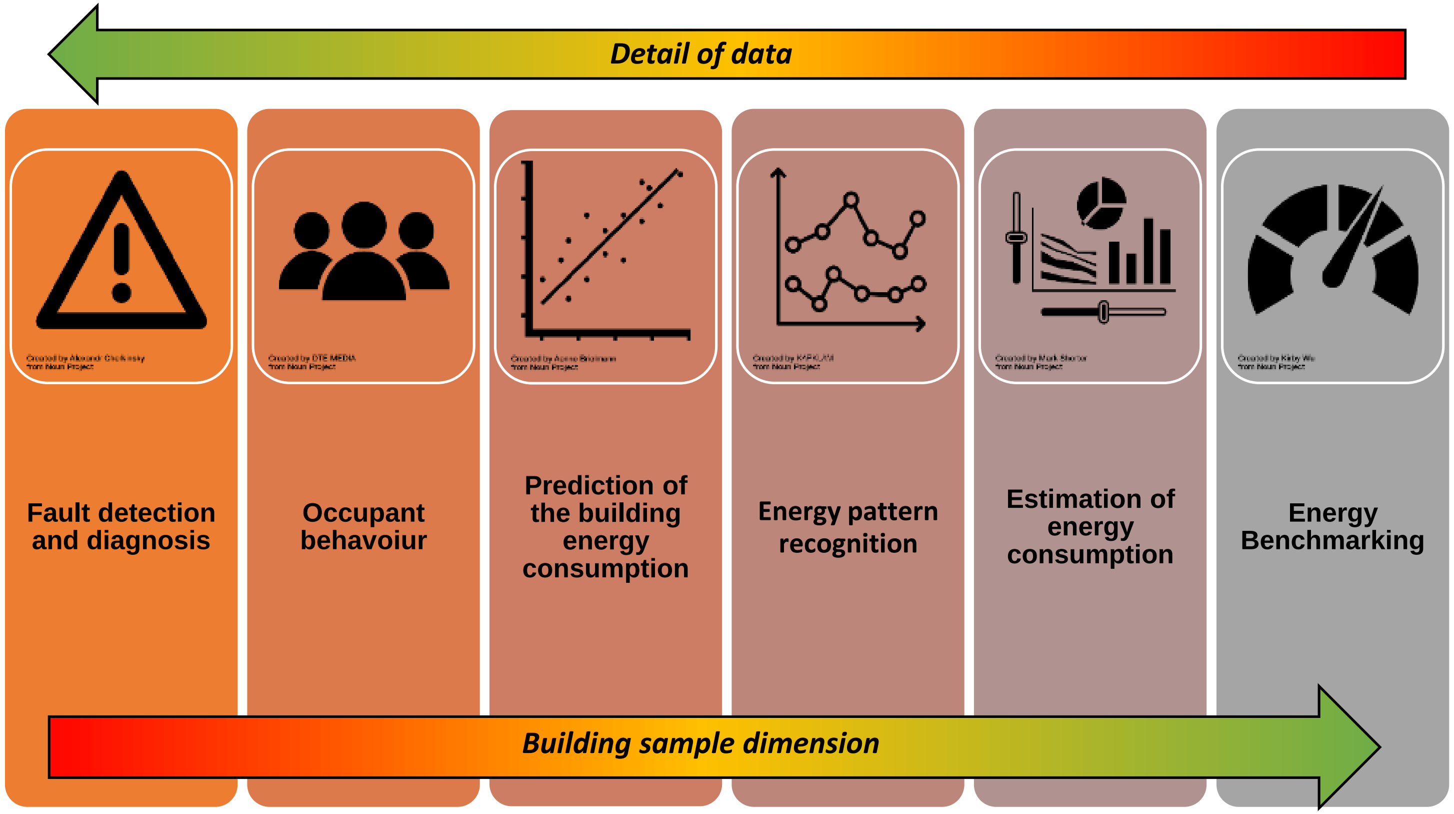
PhD student
Silvio Brandi
Topics:
Time series analysis, energy programming, analytics with R

Granted researcher
Daniele Mauro Mazzarelli
Topics:
Fault detection and diagnosis, Association rule mining

Marco Savino Piscitelli – BAEDA Research Background

Reference publications

- Capozzoli, A., **Piscitelli**, M. S., Brandi, S., Grassi, D., Chicco, G. Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings. Energy, 2018, 157, 336e352.
- Capozzoli A, **Piscitelli** M S, Brandi S. Mining typical load profiles in buildings to support energy management in the smart city context. Energy Procedia 2017; 134: 865–874. (invited paper at conference SEB17) http://linkinghub.elsevier.com/retrieve/pii/S187661021734674X
- Capozzoli A, **Piscitelli** M S, Gorrino A, Ballarini I, Corrado V. Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings. Sustain Cities Soc 2017; 35: 191–208. http://dx.doi.org/10.1016/j.scs.2017.07.016.
- Capozzoli A, **Piscitelli** M S, Neri F, Grassi D, Serale G. A novel methodology for energy performance benchmarking of buildings by means of Linear Mixed Effect Model: The case of space and DHW heating of out-patient Healthcare Centres, Applied Energy, Elsevier, 2016, 171:592-607. http://dx.doi.org/10.1016/j.apenergy.2016.03.083
- Capozzoli A, Cerquitelli T, **Piscitelli** M S. Enhancing energy efficiency in buildings through innovative data analytics technologies (book chapter). Pervasive Computing: Next Generation Platforms for Intelligent Data Collection. Elsevier 2016. Editors: Ciprian Dobre, Fatos Xhafa. http://dx.doi.org/10.1016/B978-0-12-803663-1.00011-5
- Capozzoli, A., Lauro, F., Khan, I. Fault detection analysis using data mining techniques for a cluster of smart office buildings, Expert Systems with Applications, Elsevier, 2015, 42(9), 4324–4338. https://doi.org/10.1016/j.eswa.2015.01.010



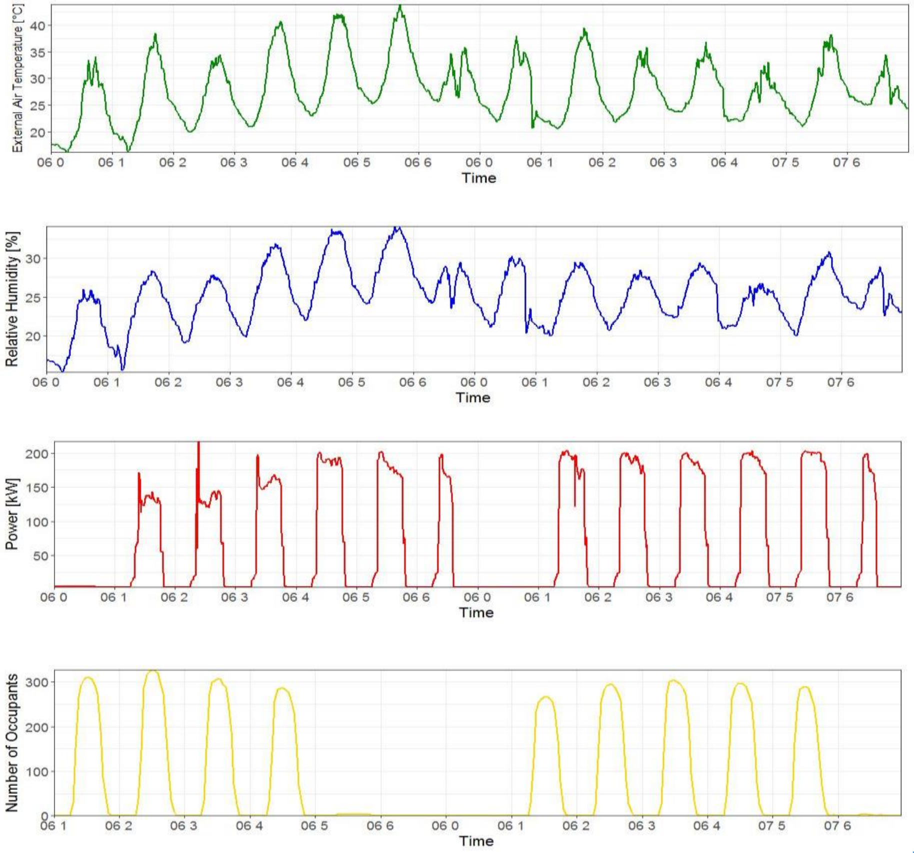
Energy profiling

The increasing implementation of ICT and EMS in the current ***paradigm of smart buildings in smart cities*** has enabled an easier availability of a huge amount of heterogeneous and complex building-related data in form of time series.

Climatic data

Operational data

User related data



What is a time series?

*A **time series** is a **series** of data points listed in **time** order. Most commonly, a **time series** is a sequence taken at successive equally spaced points in **time**.*

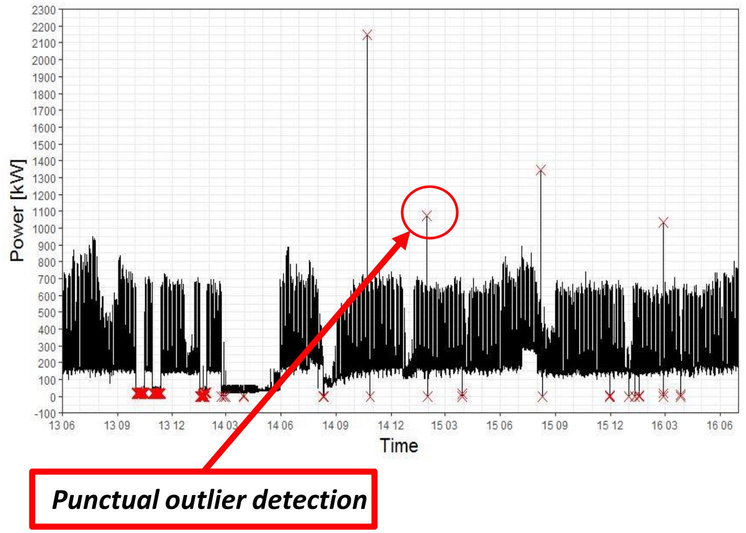
ENERGY PROFILING

The mining of ***time series data*** has recently gained high attention as a way to describe and deeply ***characterise*** typical operational patterns and trends of ***energy consumption in buildings.***

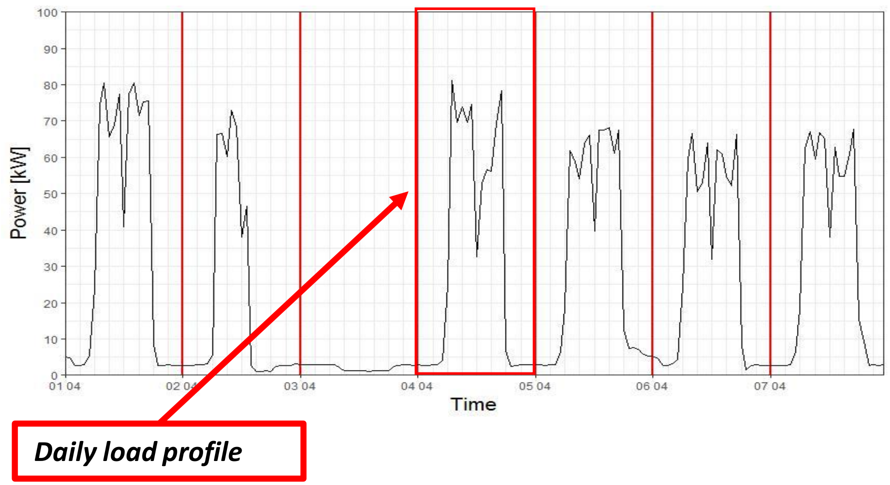
Data pre-processing

In a first step, the collected raw data in form of time series are analysed through different statistical methods to identify potential **missing values and punctual outliers** that must be replaced or removed.

DATA CLEANING



TIME-SERIES CHUNKING



In a second step, the original **time series is chunked in fixed length windows** (sub-sequences). The sub-sequences, representing the daily load profiles, are **organized into a MxN matrix** where M is the number of daily load profiles while N depends from the data granularity.

The load profiles (M by N) matrix

date	time	Power [kW]
20/09/2018	00:00	10
20/09/2018	06:00	20
20/09/2018	12:00	34
20/09/2018	18:00	20
21/09/2018	00:00	6
21/09/2018	06:00	15
21/09/2018	12:00	67
21/09/2018	18:00	30
22/09/2018	00:00	9
22/09/2018	06:00	12
22/09/2018	12:00	21
22/09/2018	18:00	9

KEY **VALUE**

N - dimension

M - dimension

date	00:00	06:00	12:00	18:00
20/09/2018	10	20	34	20
21/09/2018	6	15	67	30
22/09/2018	9	12	21	9

N – dimension = depends from the timestep of the time series
(e.g. for hourly time series N = 24)

M – dimension = depends from the number of days

Typical load profiles identification

This phase of the framework is performed at individual building/customer level and it is aimed at **identifying groups of homogenous profiles in the M by N matrix** through a data segmentation phase.

The typical profiles can be then evaluated through statistical measures (e.g. mean, median) calculated in each group of homogenous daily load profiles identified in the data segmentation phase. To this purpose, data segmentation may be performed following:

1. **Domain expert based approach.**
2. **Data mining approach by using unsupervised techniques.**
3. **Indirect clustering through data reduction methods.**

Cluster analysis

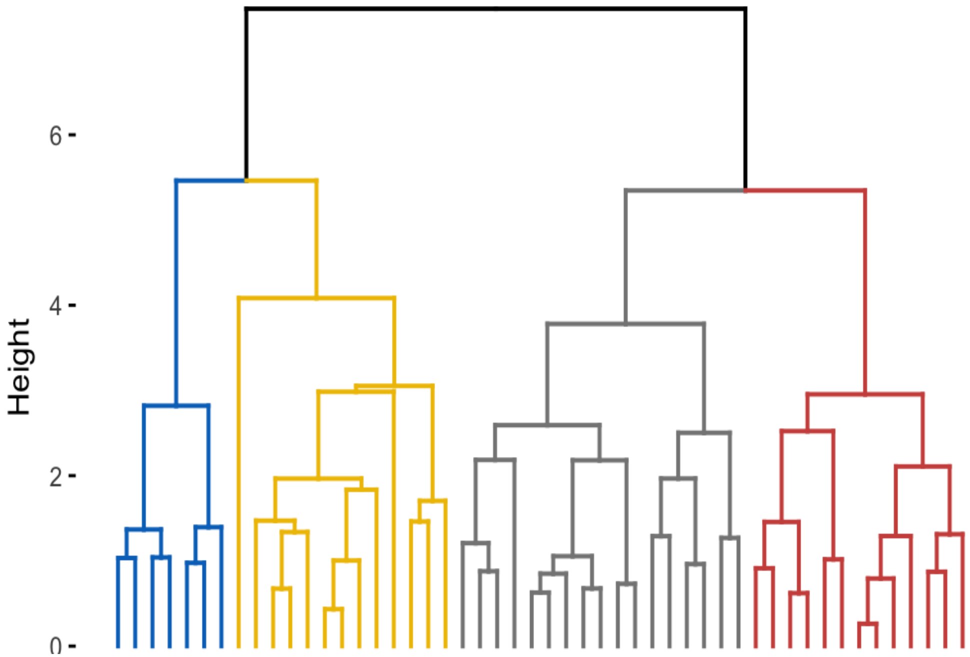
- Clustering allows to segment a set of data objects into clusters based on a concept of **similarity/proximity among data**.

- The objective of any clustering algorithm consists in **dividing a set of data composed of n multidimensional objects** {x1, . . ., xn} **into K clusters** {C1, . . ., CK}, in order to group similar objects in the same cluster and dissimilar objects into different clusters.

- The set of clusters P = {C1, . . ., CK} is referred as data partition.

Hierarchical clustering

- A set of nested clusters organized as a hierarchical bottom/up (agglomerative) or top/down (divisive) tree



Cluster analysis

MxN matrix

N - dimension

M - dimension

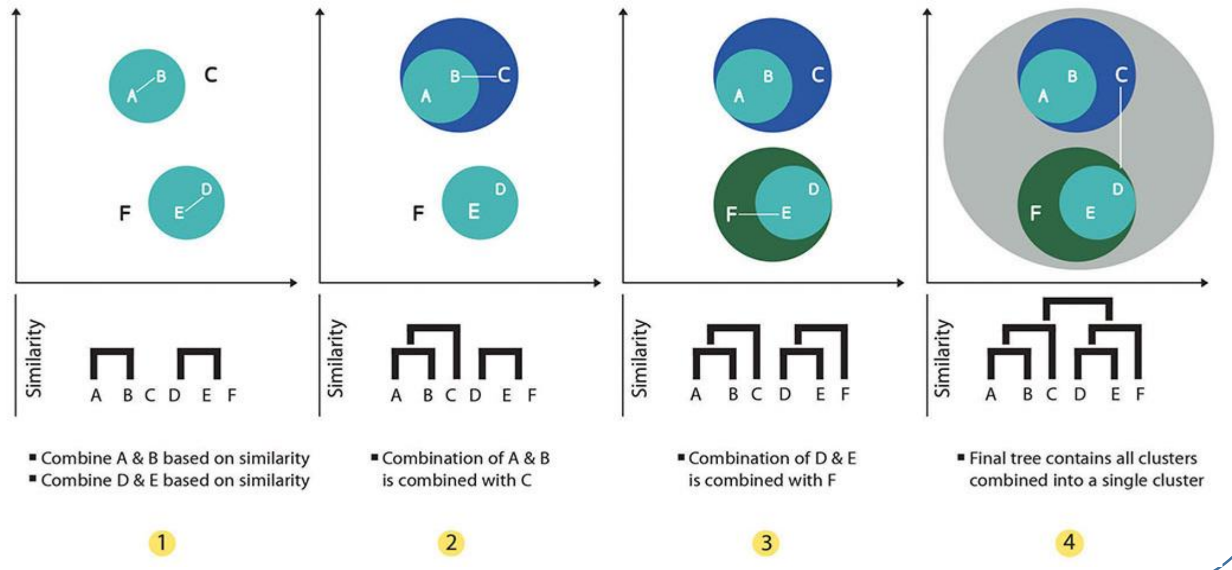
date	00:00	06:00	12:00
20/09/2018	10	20	34
21/09/2018	6	15	67
22/09/2018	9	12	21
23/09/2018	10	20	34
24/09/2018	6	15	67
25/09/2018	9	12	21

distance matrix

Euclidean distance

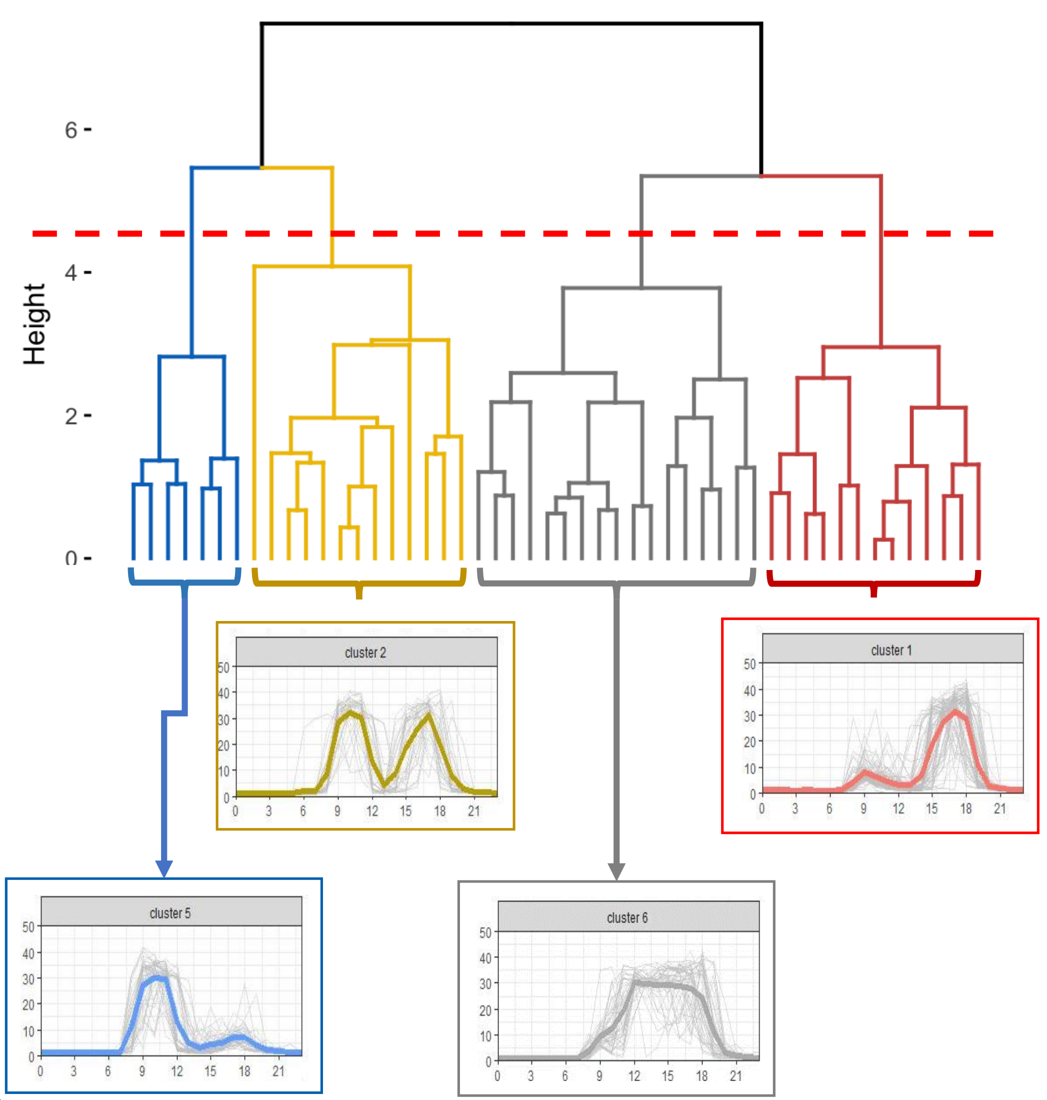
$$d_{ED}(X,Y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2}$$

$$A = \begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 & \cdots & d_{1n}^2 \\ d_{21}^2 & 0 & d_{23}^2 & \cdots & d_{2n}^2 \\ d_{31}^2 & d_{32}^2 & 0 & \cdots & d_{3n}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1}^2 & d_{n2}^2 & d_{n3}^2 & \cdots & 0 \end{bmatrix}$$



https://www.brandidea.com/hierarchicalclustering.html

Cut of the cluster dendrogram



Discovering frequent patterns in time series through unsupervised data mining techniques: the case of the energy profiling in buildings

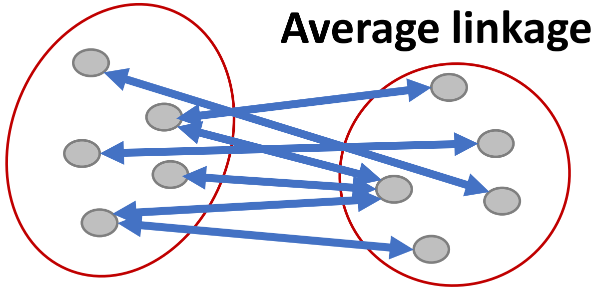
Cluster analysis hierarchical algorithms

Average linkage hierarchical clustering: In this type, two clusters whose merger has the smallest average distance between data points are merged in each step.

Complete linkage hierarchical clustering: In this type, two clusters whose merger has the smallest diameter are merged in each step.

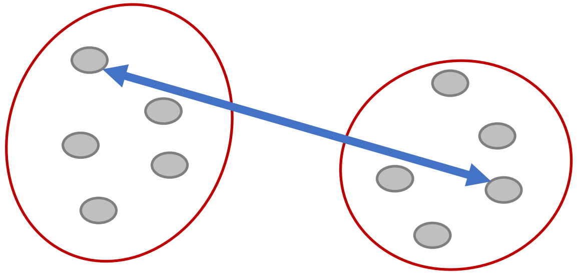
Single linkage hierarchical clustering: In this linkage type, two clusters whose two closest members have the shortest distance are merged in each step.

Average linkage



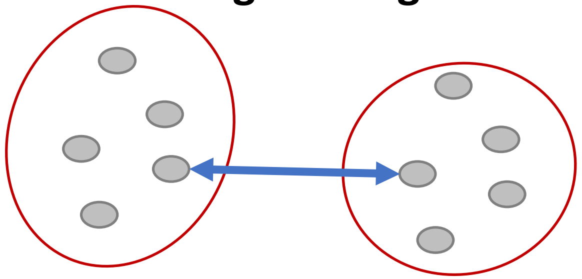
$$\mathrm{L}(r,t) = \frac{1}{n_r n_t}\sum_{i=1}^{n_r}\sum_{j=1}^{n_t} D(x_{ri}, x_{tj})$$

Complete linkage



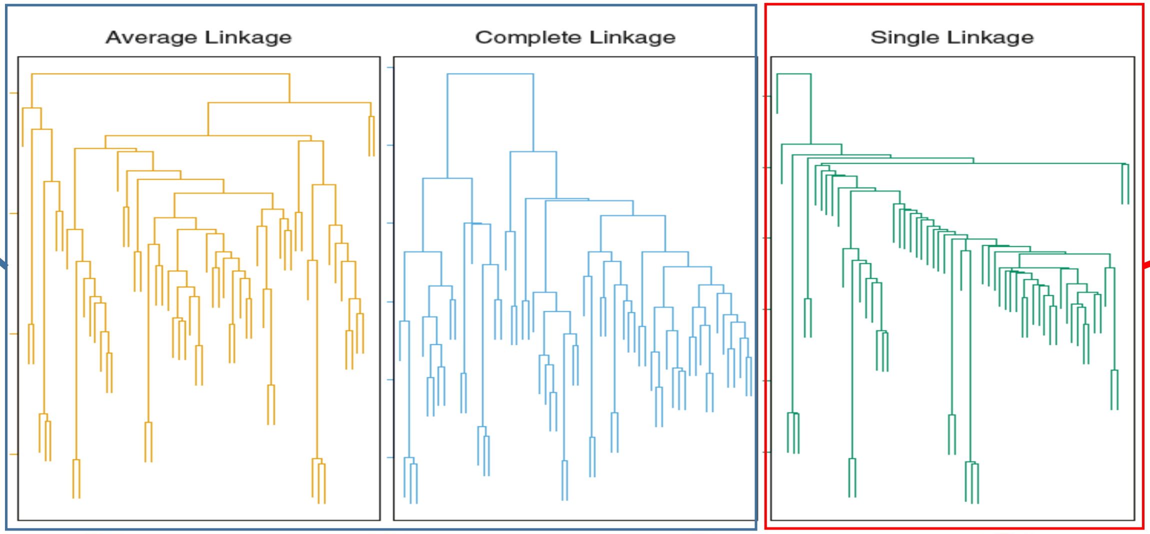
$$\mathrm{L}(r,t) = \max(D(x_{ri}, x_{tj})$$

Single linkage



$$\mathrm{L}(r,t) = \min(D(x_{ri}, x_{tj})$$

Average and Complete produce more balanced dendrograms.



Single is more sensitive to outliers.

https://www.brandidea.com/hierarchicalclustering.html

Cluster analysis – optimal number of clusters

Select the validation metrics

	Name of the index in **NbClust**	Optimal number of clusters
1.	"ch" (Calinski and Harabasz 1974)	Maximum value of the index
2.	"duda" (Duda and Hart 1973)	Smallest number of clusters such that index > criticalValue
3.	"pseudot2" (Duda and Hart 1973)	Smallest number of clusters such that index < criticalValue
4.	"cindex" (Hubert and Levin 1976)	Minimum value of the index
5.	"gamma" (Baker and Hubert 1975)	Maximum value of the index
6.	"beale" (Beale 1969)	Number of clusters such that critical value >= alpha
7.	"ccc" (Sarle 1983)	Maximum value of the index
8.	"ptbiserial" (Milligan 1980, 1981)	Maximum value of the index
9.	"gplus" (Rohlf 1974; Milligan 1981)	Minimum value of the index
10.	"db" (Davies and Bouldin 1979)	Minimum value of the index
11.	"frey" (Frey and Van Groenewoud 1972)	Cluster level before index value < 1.00
12.	"hartigan" (Hartigan 1975)	Maximum difference between hierarchy levels of the index
13.	"tau" (Rohlf 1974; Milligan 1981)	Maximum value of the index
14.	"ratkowsky" (Ratkowsky and Lance 1978)	Maximum value of the index
15.	"scott" (Scott and Symons 1971)	Maximum difference between hierarchy levels of the index
16.	"marriot" (Marriot 1971)	Max. value of second differences between levels of the index
17.	"ball" (Ball and Hall 1965)	Maximum difference between hierarchy levels of the index
18.	"trcovw" (Milligan and Cooper 1985)	Maximum difference between hierarchy levels of the index
19.	"tracew" (Milligan and Cooper 1985)	Max. value of second differences between levels
20.	"friedman" (Friedman and Rubin 1967)	Maximum difference between hierarchy levels of the index
21.	"mcclain" (McClain and Rao 1975)	Minimum value of the index
22.	"rubin" (Friedman and Rubin 1967)	Minimum value of second differences between levels
23.	"kl" (Krzanowski and Lai 1988)	Maximum value of the index
24.	"silhouette" (Rousseeuw 1987)	Maximum value of the index
25.	"gap" (Tibshirani *et al.* 2001)	Smallest number of clusters such that criticalValue >= 0
26.	"dindex" (Lebart *et al.* 2000)	Graphical method
27.	"dunn" (Dunn 1974)	Maximum value of the index
28.	"hubert" (Hubert and Arabie 1985)	Graphical method
29.	"sdindex" (Halkidi *et al.* 2000)	Minimum value of the index
30.	"sdbw" (Halkidi and Vazirgiannis 2001)	Minimum value of the index

Table 2: Overview of the indices implemented in the **NbClust** package.

Select the search space (e.g. from 2 to 15 clusters)

the 70% of cluster analyses result in less than 7 clusters

Select the optimal number according the majority rule

```
*******************************************************************
* Among all indices:
* 2 proposed 2 as the best number of clusters
* 15 proposed 3 as the best number of clusters
* 6 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 3 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
```

Malika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1-36. URL http://www.jstatsoft.org/v61/i06/.

Supervised classification – Decision trees

The **cluster label** is defined **as a categorical dependent variable** which can be predicted with a classification model using additional attributes for the **supervised classification process**.

Explanatory variables

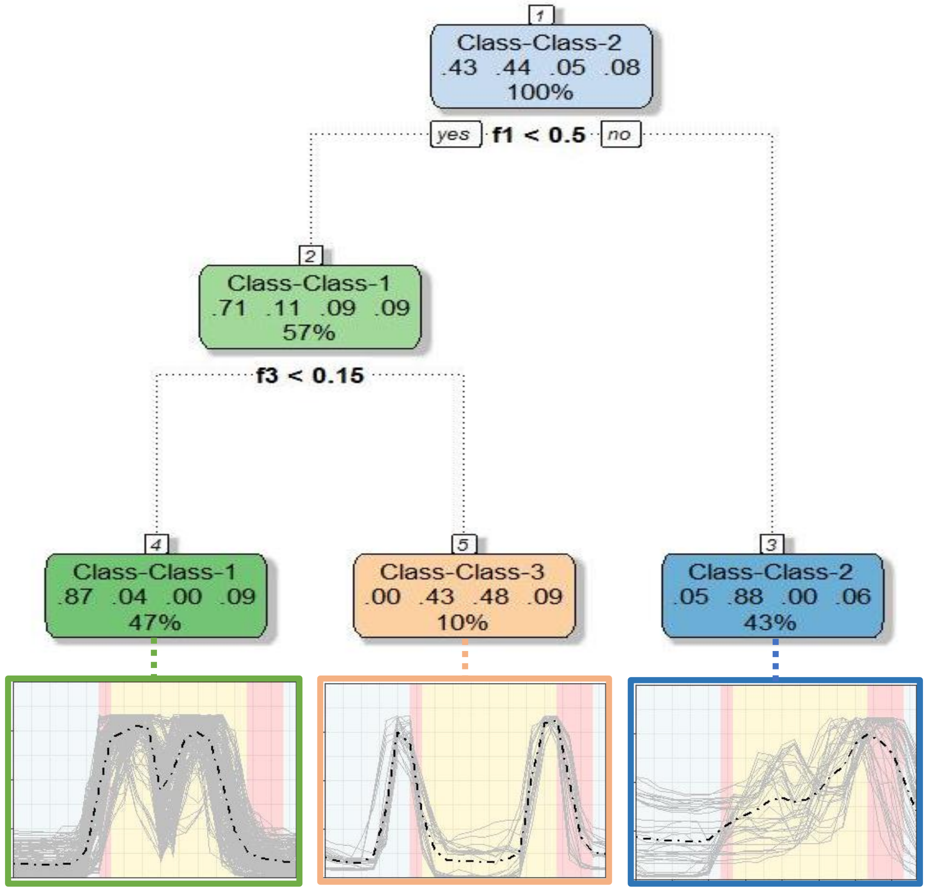
- **Time variables**
- **Energy consumption influencing variables**

$daily\ average\ solar\ radiation$

$type\ of\ the\ day\ (working, no-working)$

$daily\ average\ external\ temperature$

$day\ of\ the\ week$



Classification and regression trees

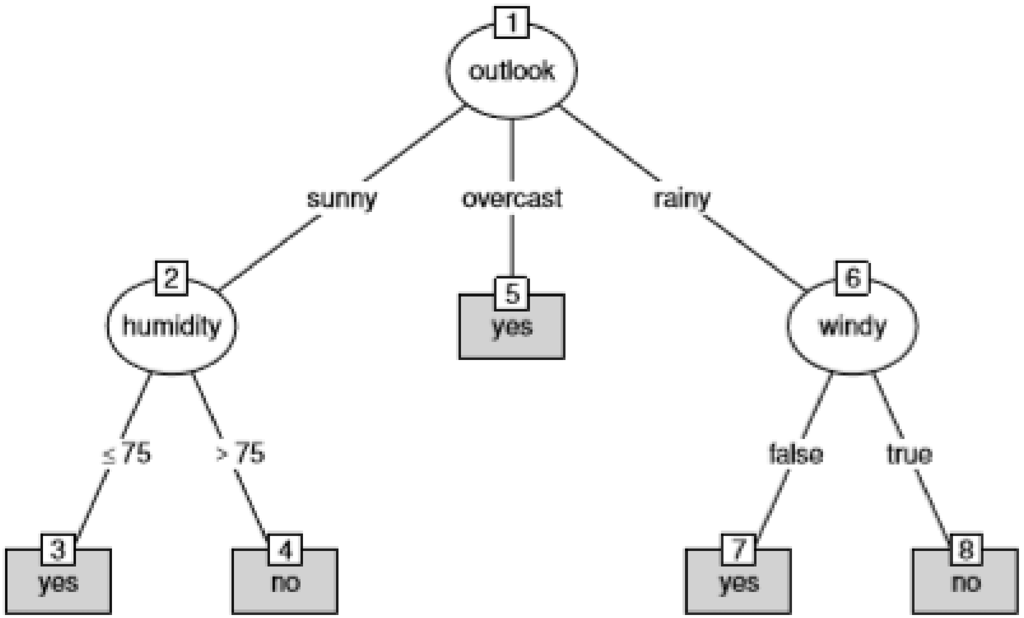
- The task of classification consists in develop a model capable to assign objects to one of different predefined class, and to predict for a new statistical object its class membership accordingly.

- The objective of a classification model consists in learning a function or a set of rules, which allows to **predict** for a new unlabeled statistical object its **class membership** and provide a **description of the data features** that characterize **objects with the same label**.

Decision trees

- Starting from the **root node**, at each node of the tree model, the data are successively splitted.
- At each split the model identify which data feature, and threshold value, better discriminate labels in the corresponding subset of data according to impurity measures.
- Tree models where the output variable takes class label are called **classification trees**, while Decision trees where the output variable takes continuous values are called **regression trees**

Classification Tree



Torsten Hothorn, Achim Zeileis (2015). partykit: A Modular Toolkit for Recursive Partytioning in R. Journal of Machine Learning Research, 16, 3905-3909. URL http://jmlr.org/papers/v16/hothorn15a.html

1. Data pre-processing
   - Replace missing values
   - Construction of MxN matrix
   - Data normalization

2. Load profile clustering
   - Construction of the distance matrix
   - Hierarchical clustering analysis
   - Evaluation of cluster centroids

3. Cluster label classification
   - Enrichment of the dataset with predictive variables
   - Development of a classification tree

Let’s code!

RStudio

Discovering frequent patterns in time series through unsupervised data mining techniques: the case of the energy profiling in buildings

Marco Savino Piscitelli

marco.piscitelli@polito.it

https://www.linkedin.com/in/marco-savino-piscitelli-387761b7/

https://www.researchgate.net/profile/Marco_Piscitelli2

PyData

20/09/2018