Modelling energy efficiency of buildings based on open-data

Tania CERQUITELLI
Department of Control and Computer engineering, Politecnico di Torino, Italy

Multidisciplinary research team

Professors of Politecnico di Torino with orthogonal multidisciplinary skills:

Prof. Tania Cerquitelli (DAUIN) – Principal Investigator
Prof. Elena Baralis (DAUIN)
Prof. Marco Mellia (DET)
Prof. Alfonso Capozzoli (DENERG)

Research fellows:

Evelina Di Corso (DAUIN)
Stefano Proto (DAUIN)
Daniele Mauro Mazzarelli (DAUIN)

Edison researchers:

Ing. Silvia Casagrande
Ing. Martina Tamburini

Main research objective

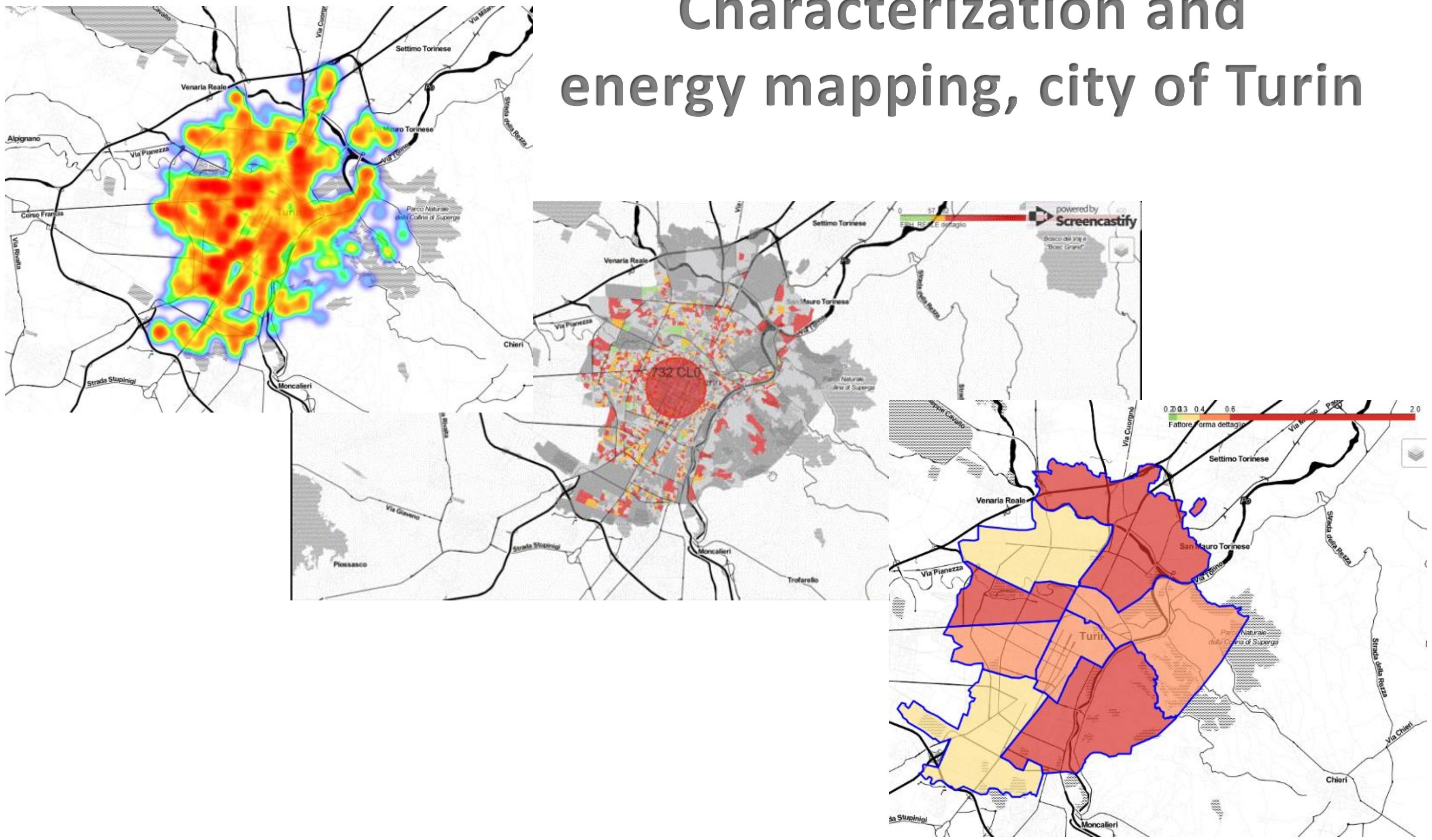
ENERGY DATA

OPEN DATA

Value for different stakeholders

Support and improve decisional processes

Characterization and energy mapping, city of Turin



EDISON
EDF GROUP

POLITECNICO DI TORINO
SmartData@PoliTO
SmartData

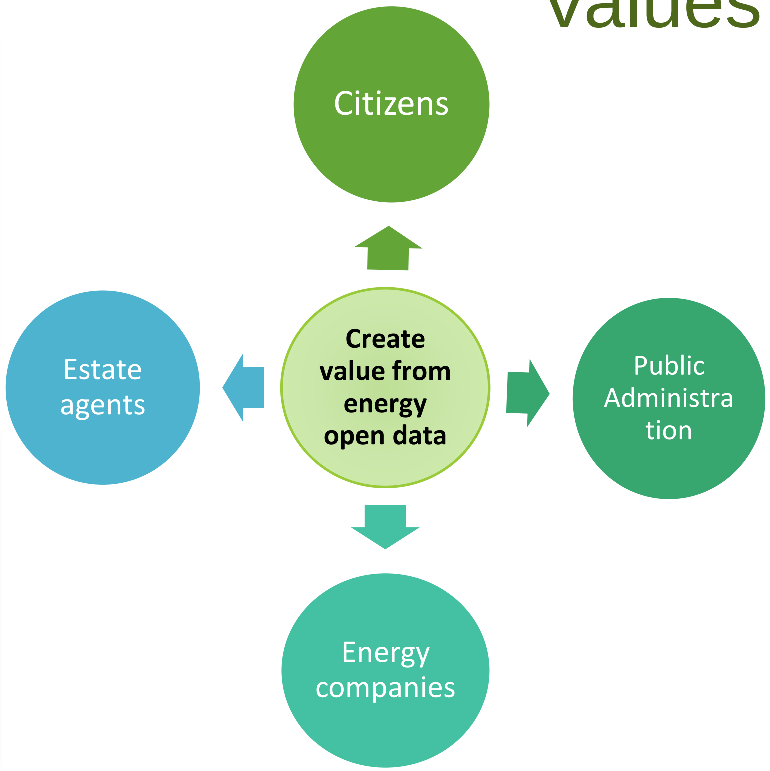
Main research objective

ENERGY DATA

OPEN DATA

Value for different stakeholders

Support and improve decisional processes



Values for the stakeholders

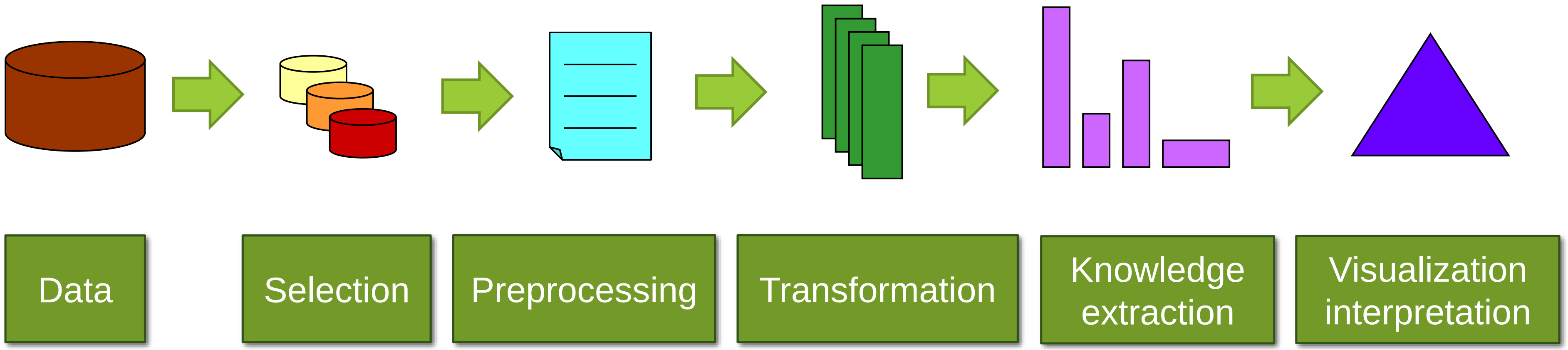
- ✓ **Mapping the energy demand** of buildings at neighborhood and city level
- ✓ **Characterization of metropolitan areas** with respect to energy-efficiency parameters
- ✓ **Targeted incentive policies**
- ✓ Energy planning
- ✓ Development of **more accurate benchmark models**
- ✓ Targeted **promotional offers**

EDISON EDF GROUP

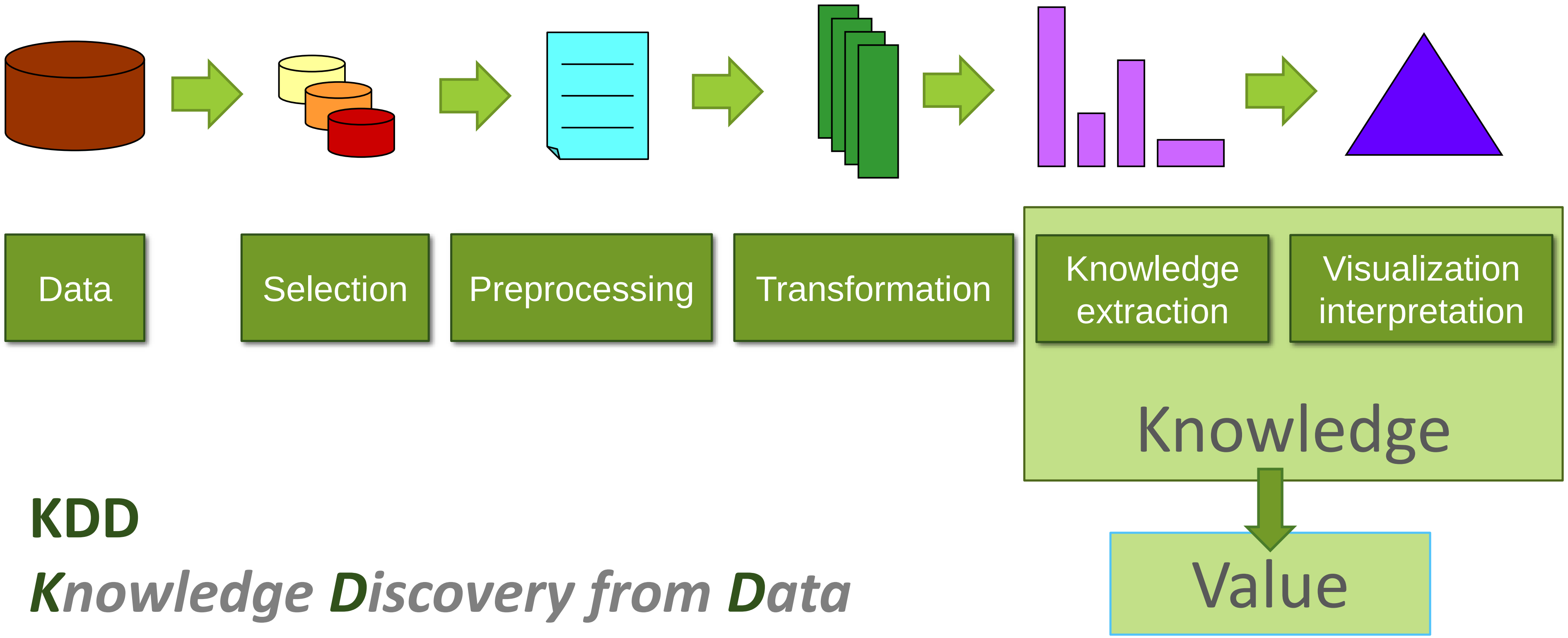
POLITECNICO DI TORINO

SmartData@PoliTO

Knowledge extraction process



Knowledge extraction process



KDD

Knowledge Discovery from Data

KDD from energy data: two key roles

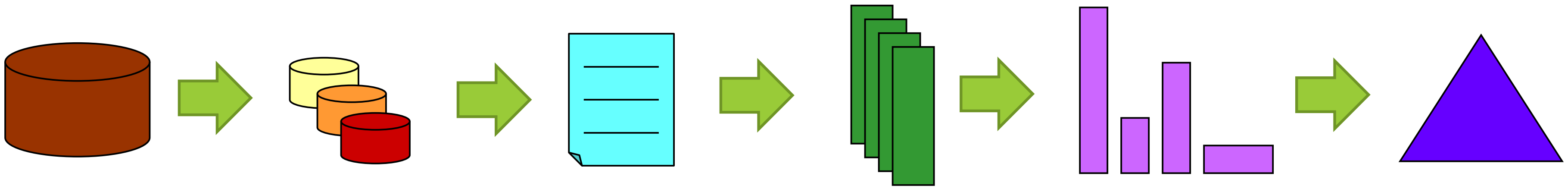
DATA SCIENTIST

- Design **innovative and efficient algorithms**
- Select the **optimal techniques** to address the challenges of the analysis
- Identify the best **trade-off** between knowledge quality and execution time

ENERGY SCIENTIST

- Support **the data pre-processing** phase
- **Assess** extracted **knowledge**
- Strong involvement in the algorithm definition phase, which should **respect/include physical laws** and correctly **model physical events**

Knowledge extraction process

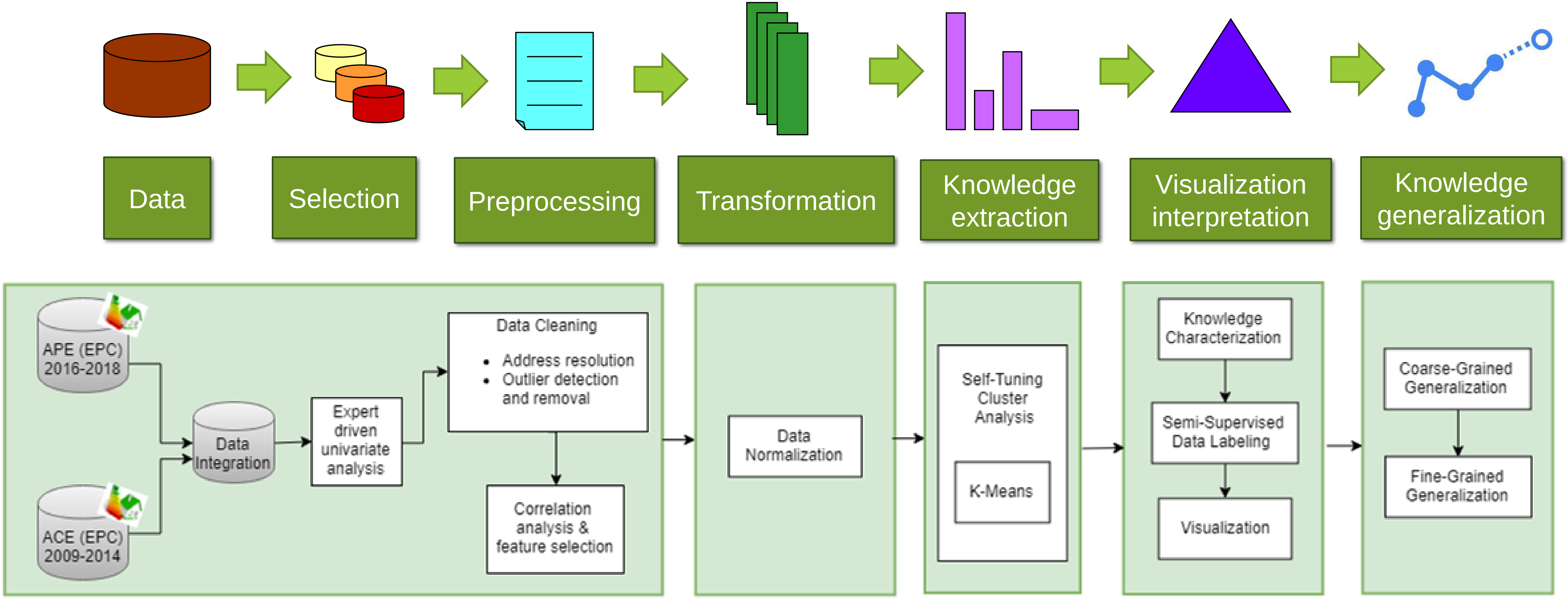


Data → Selection → Preprocessing → Transformation → Knowledge extraction → Visualization interpretation

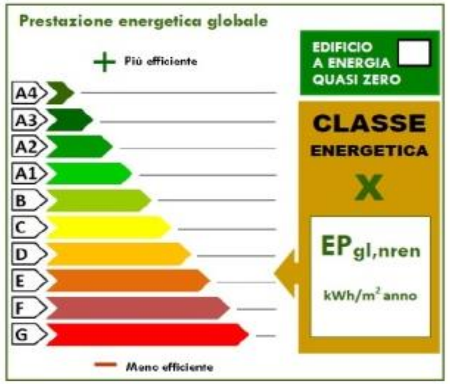
Innovations in the data analytics process

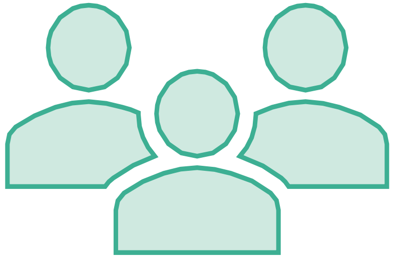
- **Tailor** the **analytic** steps to the different key aspects of **energy data**
- **Automate** the data analytics workflow to **reduce the manual user intervention**
- Translate the domain-expert knowledge into **automated procedures**
- **Generalize the extracted knowledge**
- Design **informative dashboards** to support the translation of the extracted knowledge into effective actions

Knowledge extraction process from EPCs



Open data: Energy Certificate of Buildings





Energy analysis of the building

Walling and window characteristics

Geometric features of the building

Hot water production

Environment cooling and heating

Type of plant

Renewable-energy production systems

Energy certificate officer

Qualified technicians granting APE certificates

Use of specific software (this information is not available in open data)

Building purchases

Lease agreements

Interventions to improve the building energy efficiency

Case study: EPCs in Piedmont Region

Open data available on the Sistema Piemonte service system *
Each APE is characterized by **175 attributes**, both categorical and numerical

Real building

- **Thermo-physical** characteristics (e.g., Average U-value of the vertical opaque envelope/Average U-value of the windows)
- **Geometric** features (e.g. Heated volume, Heat transfer surface, Aspect ratio)
- **Plant** characteristics (e.g. Efficiencies of the heating plant subsystems)
- **Energy** performance (e.g. Energy demands for different energy services: heating, cooling, ACS and lighting)

Reference building

- Thermo-physical characteristics
- Geometric features
- Plant characteristics
- Energy performance

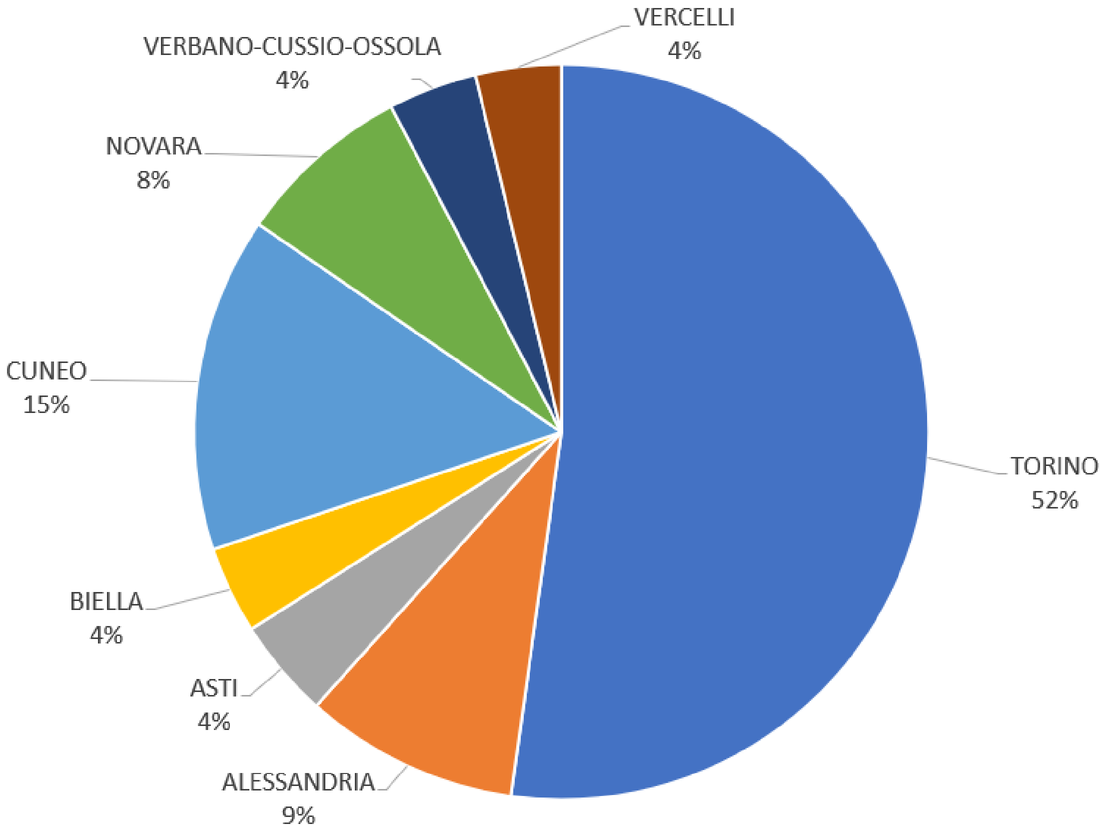
Recommendations

- Possible **actions** to improve energy performance of the building

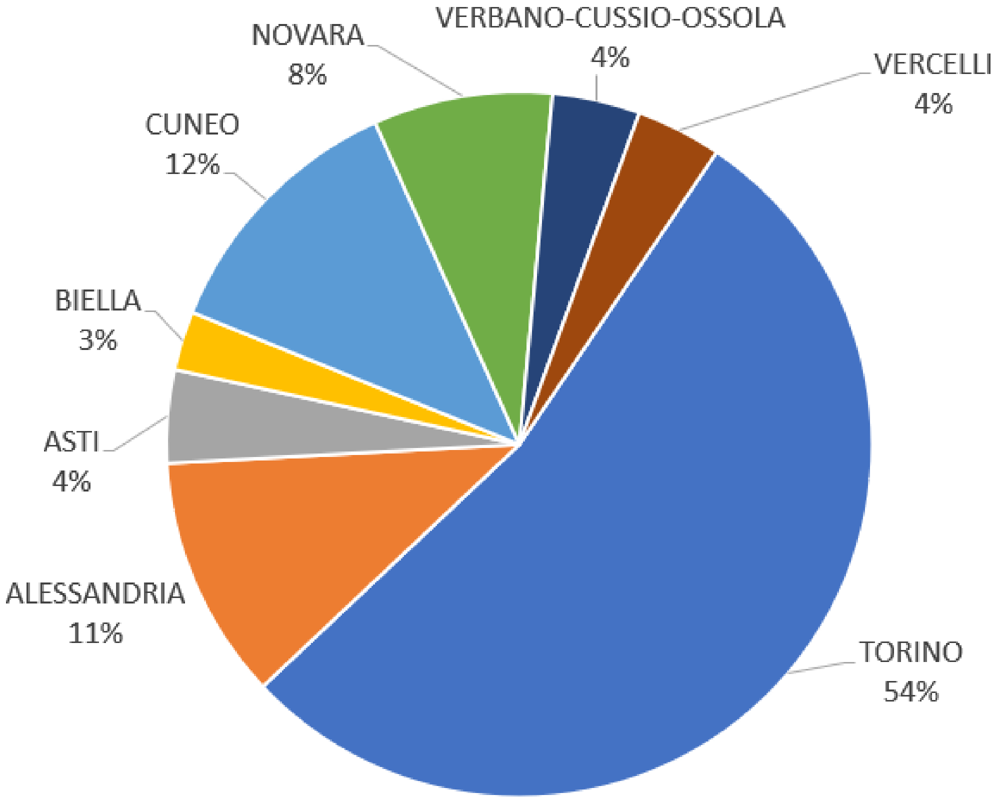
* http://www.sistemapiemonte.it/cms/privati/ambiente-e-energia/servizi/856-sistema-informativo-per-le-prestazioni-energetiche-degli-edifici-sipee

EPCs in Piedmont Region: 2 data sources

Distribution of the number of EPCs by **province**

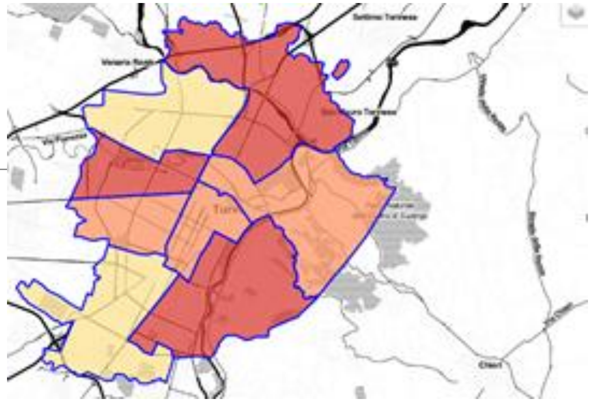


*Reference period **2009 – 2014***
EPC no. 190,124

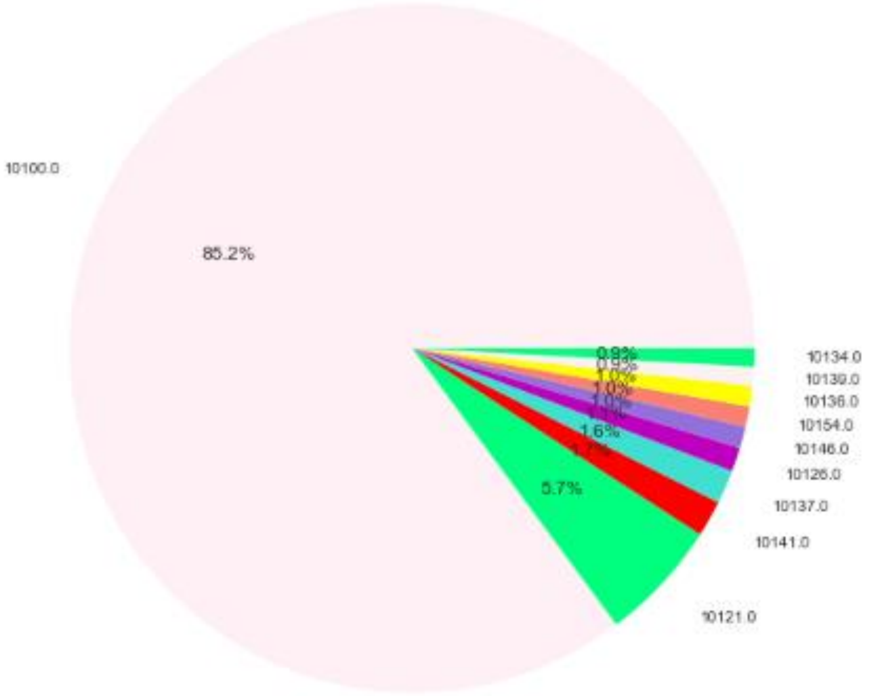


*Reference period **2015 – 06/2018***
EPC no. 78,733

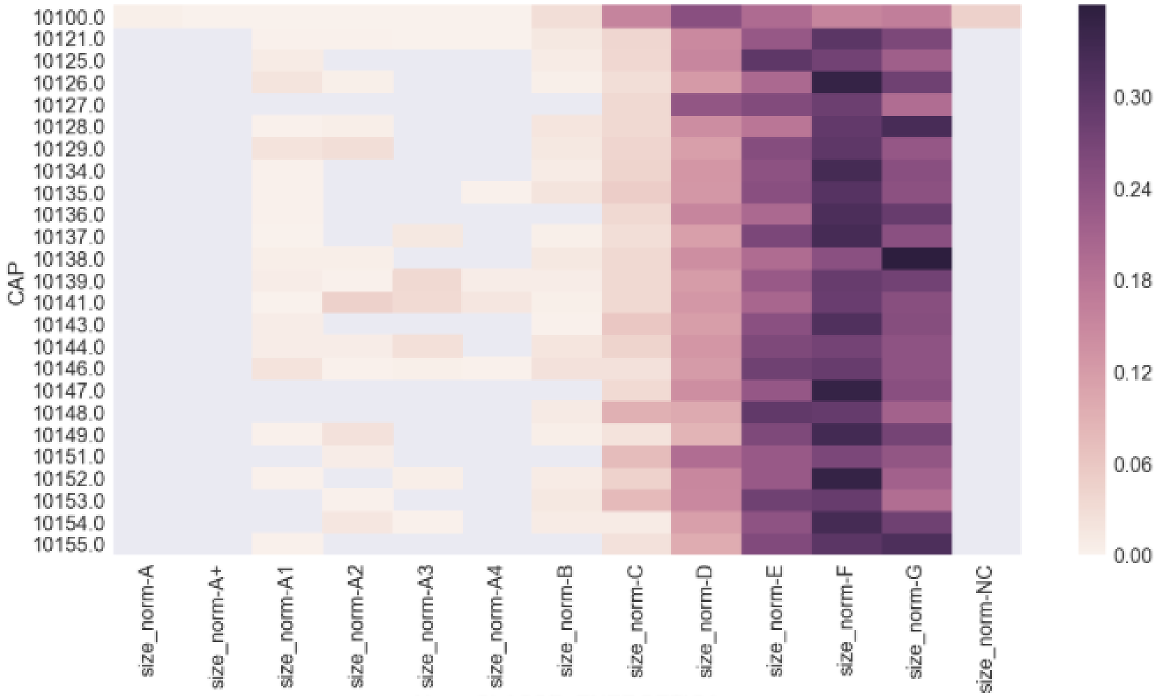
Case study: Turin



- *The city of has been selected for the variability and cardinality of EPCs in the dataset*
- *The number of EPCs is 47,623*

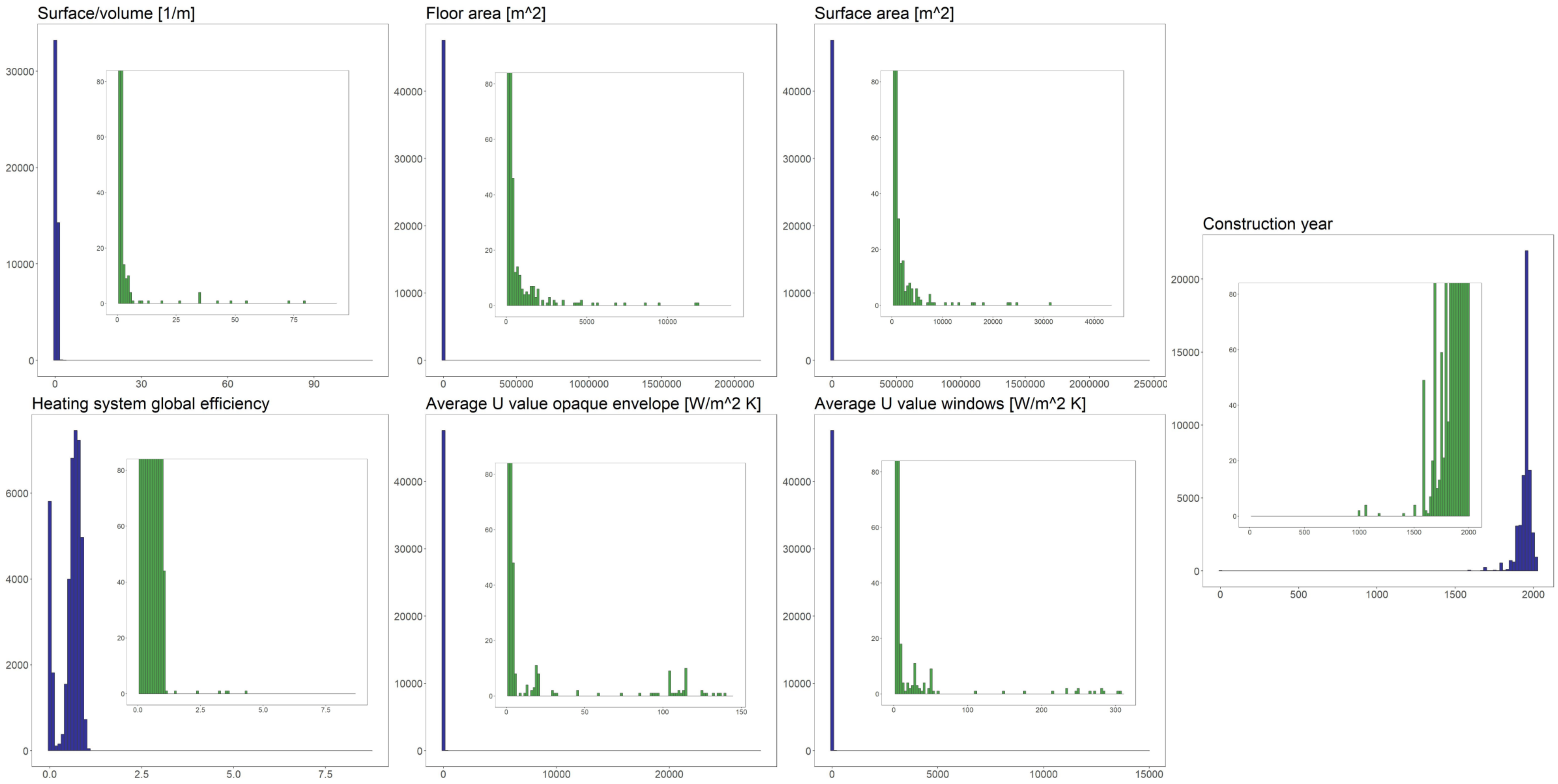


Top 15 ZIP code in Turin

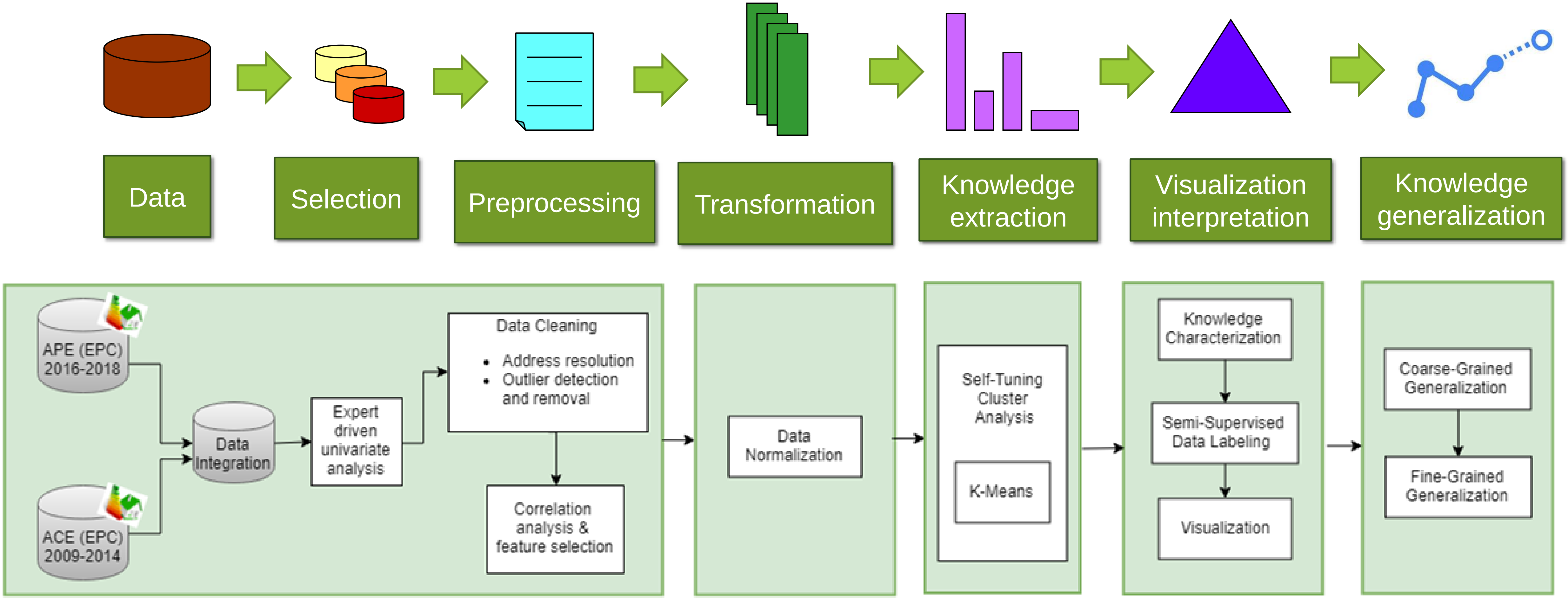


EPC# Normalized with respect to ZIP codes (only to 15 ZIP code)

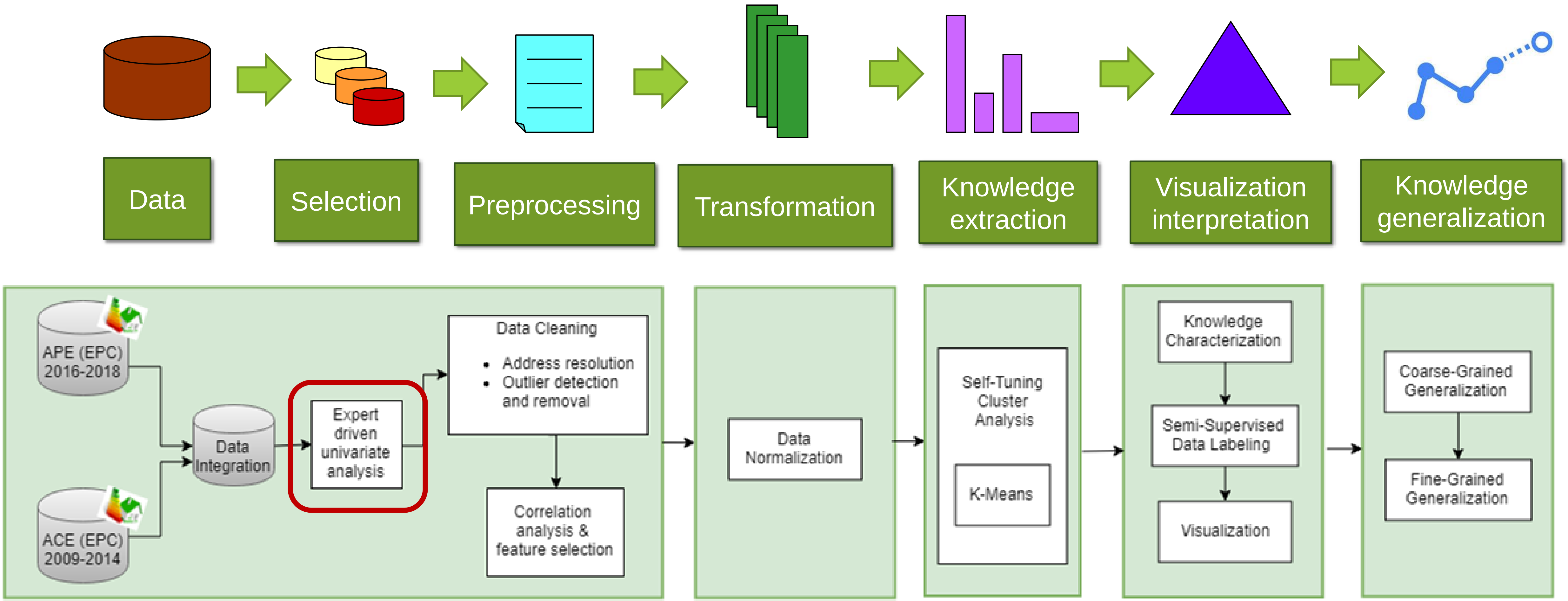
Data characterization: EPCs in Turin



Knowledge extraction process from EPCs



Knowledge extraction process from EPCs



Expert-driven univariate analysis

E1 (1) buildings used as permanent residence.

Identification of the most important variables

- Normalized Primary heating energy consumption
- Aspect Ratio
- Surface area
- Floor area
- Average U-value of the vertical opaque envelope
- Average U-value of the windows
- Heating system global efficiency
- Construction year

Expert-driven univariate analysis

E1 (1) dwellings used as permanent residence.

Identification of the most important variables → Identification of the validity ranges for each variable

- Normalized Primary heating energy consumption
- Aspect Ratio
- Surface area
- Floor area
- Average U-value of the vertical opaque envelope
- Average U-value of the windows
- Heating system global efficiency
- Construction year

Semi-supervised outlier detection

- **Definition of acceptability ranges**
- **Univariate outlier detection** based on **gESD method** needs as input parameter the upper-bound of potential outliers
- **Analysis of data distribution through Boxplot:** visualization of a data distribution through its quartiles

Attribute	**Units**	**min**	**max**
Normalised Primary heating energy consumption	$[KWh/m^2]$	0	682
Surface/Volume (S/V) ratio	$[m^{-1}]$	0.1	2
Surface area	$[m^2]$	24.9	880
Floor area	$[m^2]$	21.5	296
Average U-value of the vertical opaque envelope (U-value opaque)	$[W/m^2K]$	0.15	3
Average U-value of the windows (U-value transparent)	$[W/m^2K]$	0.9	7
Heating system global efficiency[4]	-	0.3	1.06
Construction year	-	1700	2018

gESD = generalized Extreme Studentized Deviate

Expert-driven univariate analysis

E1 (1) dwellings used as permanent residence.

Identification of the most important variables

- Normalized Primary heating energy consumption
- Aspect Ratio
- Surface area
- Floor area
- Average U-value of the vertical opaque envelope
- Average U-value of the windows
- Heating system global efficiency
- Construction year

Identification of the validity ranges for each variable

EPCs characterized by admissible values

~ 63%

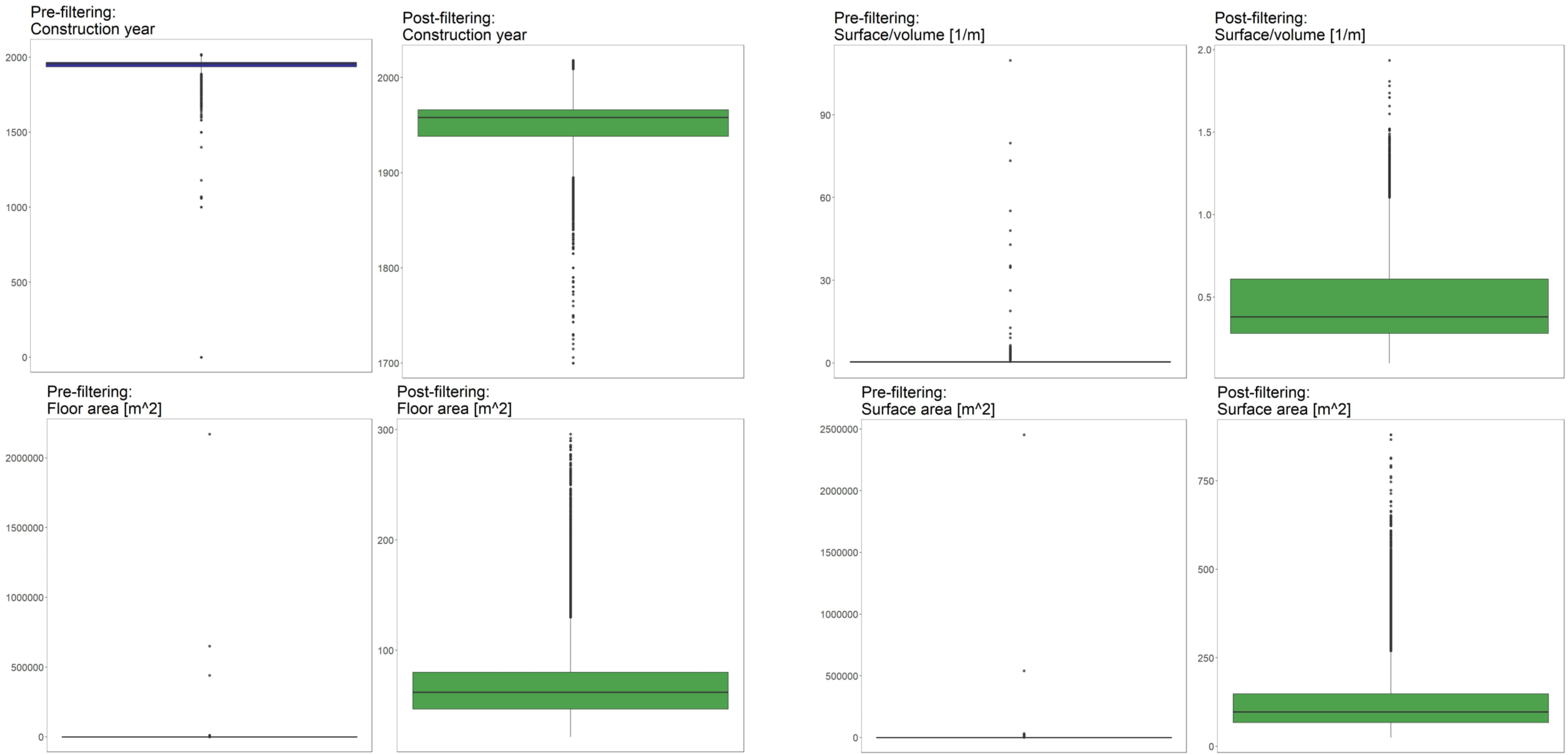
EPCs that require further investigation

Semi-supervised outlier detection

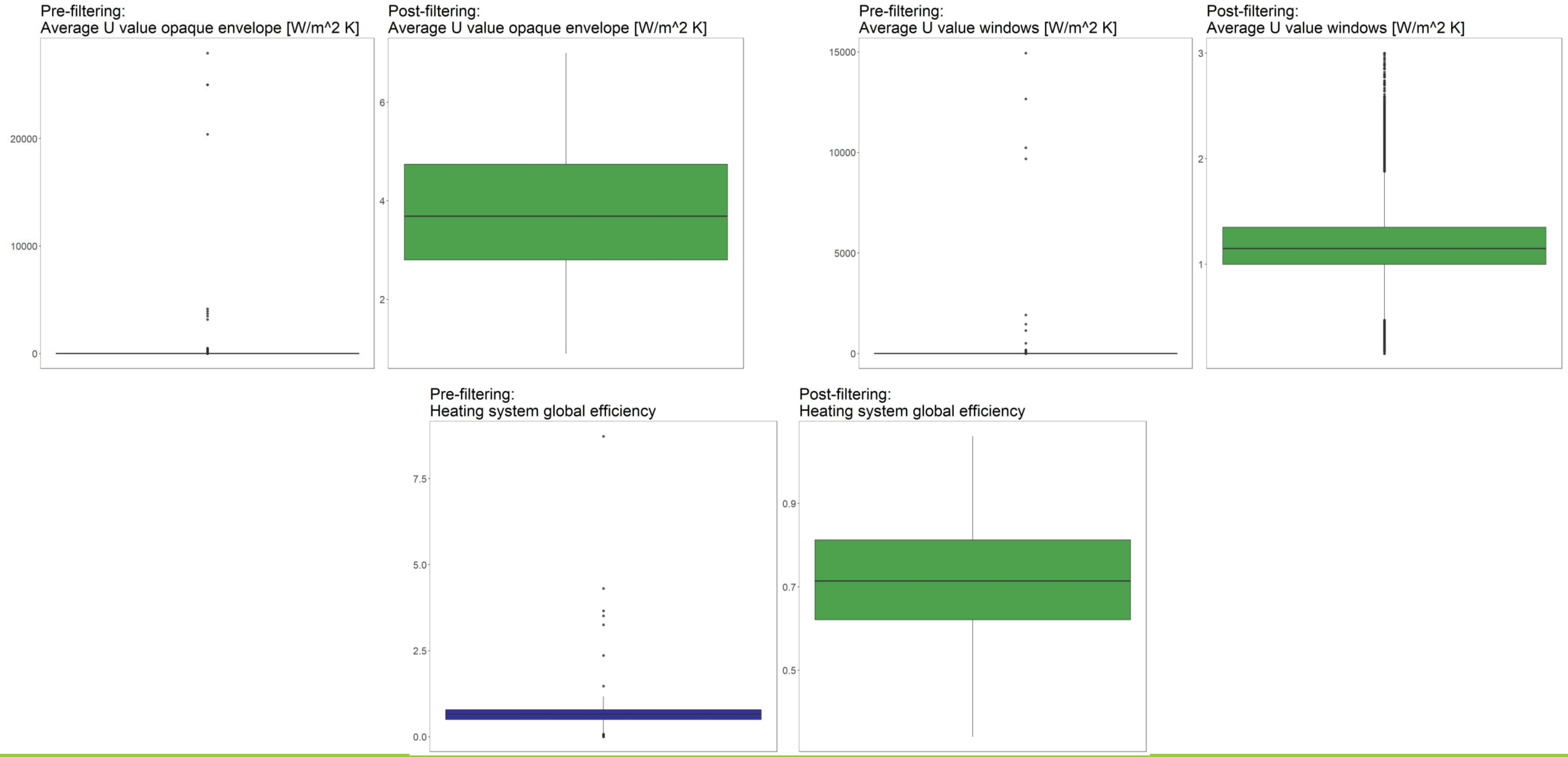
- **Definition of acceptability ranges**
- **Univariate outlier detection** based on **gESD method** needs as input parameter the upper-bound of potential outliers
- **Analysis of data distribution through Boxplot:** visualization of a data distribution through its quartiles

gESD = generalized Extreme Studentized Deviate

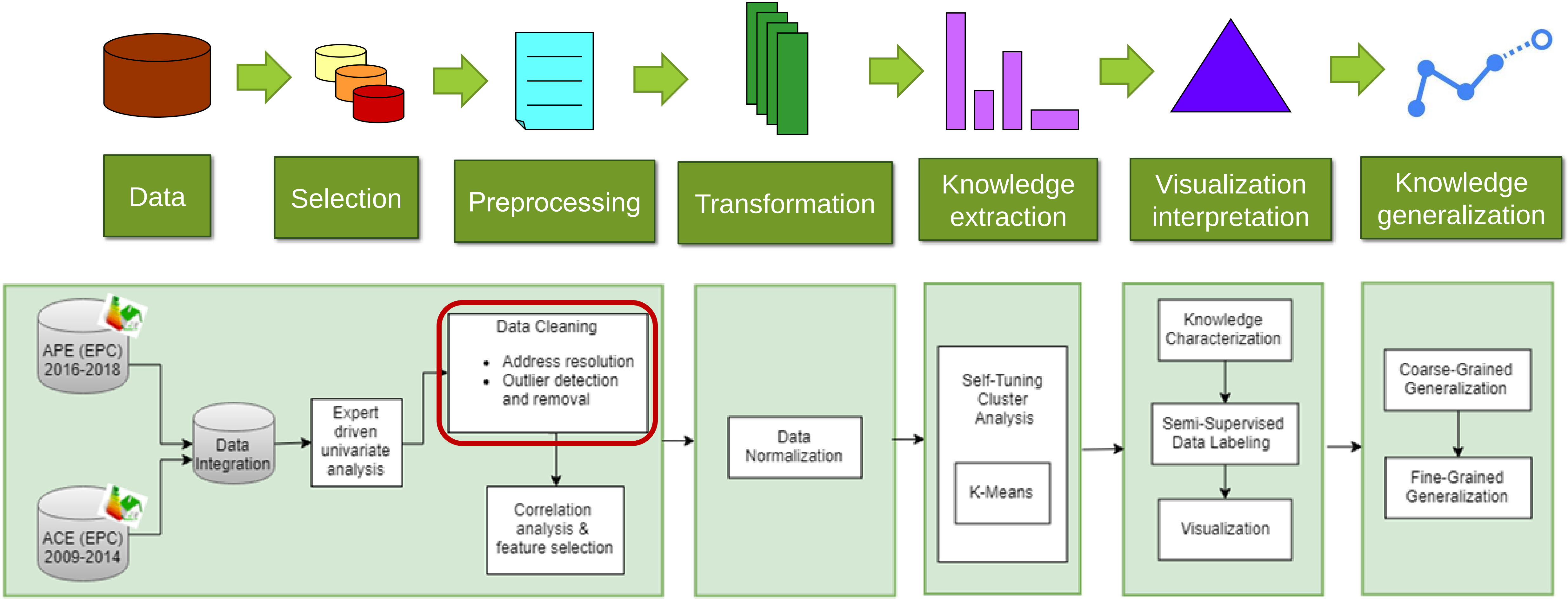
Effects of the acceptability ranges



Effects of the acceptability ranges



Preprocessing-Correlation Analysis



Data cleaning: address resolution

EPCs with invalid address format

- Typing errors
- Incorrectly-coded characters
- 31.6% of the addresses have a generic 10100 CAP
- Wrong longitude and longitude coordinates

Adopted solution

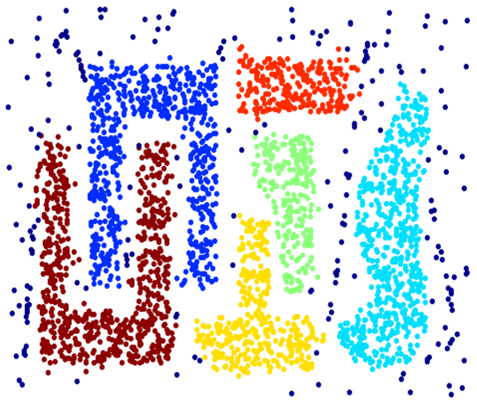
- Addresses in the DB have been **compared** to those stored in the **Turin road list** (from ***Geoportale Comune di Torino***[1])
- **Levenshtein** distance to compute the similarity index between the addresses reported in the APE DB and the reference DB.
  - If the address has been **resolved**, the CAP and the coordinates are saved in our DB eliminating inconsistencies
  - If the address has **not** been **resolved**, the CAP and coordinates are obtained through the Google[2] geocoding API
- More than 99% of the addresses have been solved

1 https://developers.google.com/maps/documentation/geocoding/intro
2 http://geoportale.comune.torino.it/web/

Outlier detection: multivariate analysis

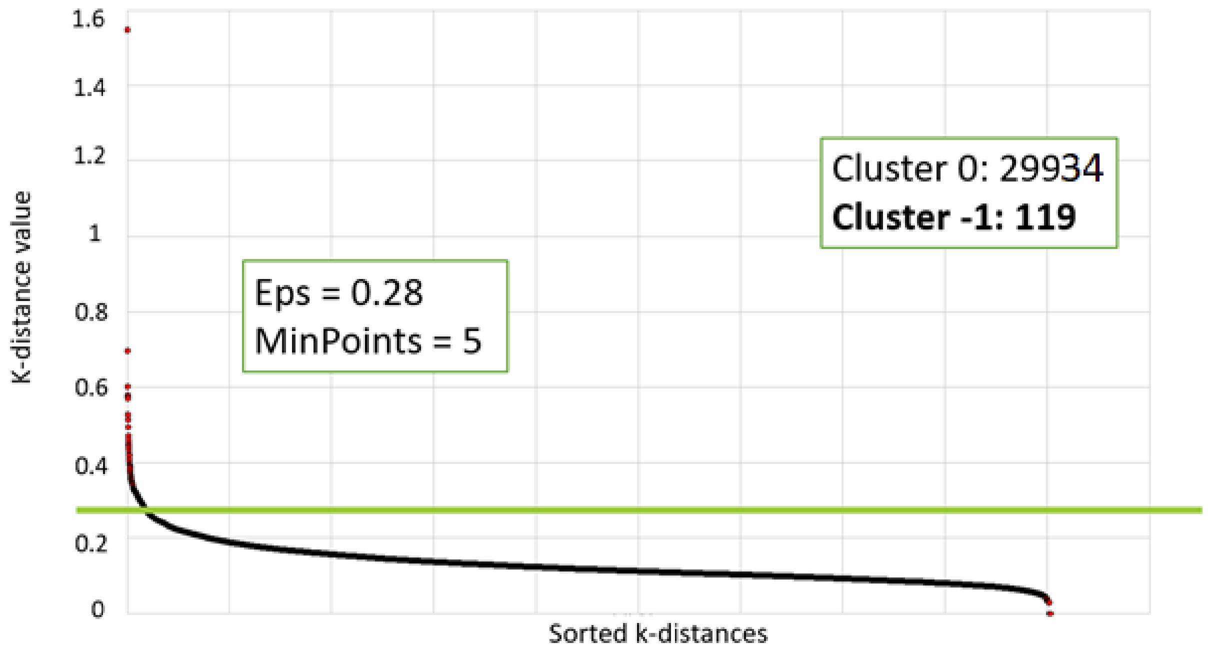
Density-based clustering algorithm: **DBScan**

- Splits the database in parts characterized by different densities (dense and sparse)
- **Density** is defined by two parameters (i.e., Eps, MinPoints), that are difficult to set
- Self-tuning strategy based on k-distances plot
  - sorted distance of every point to its kth nearest neighbor

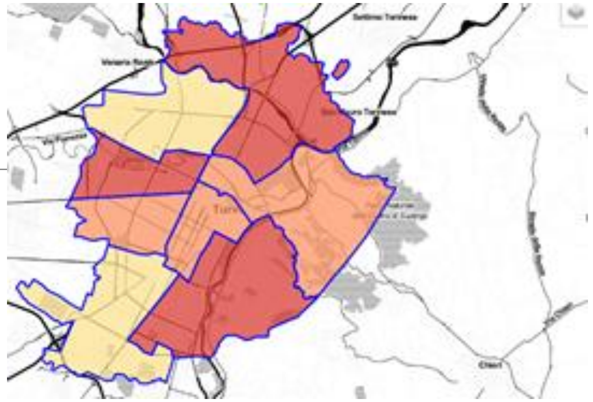


Clustering with DBScan

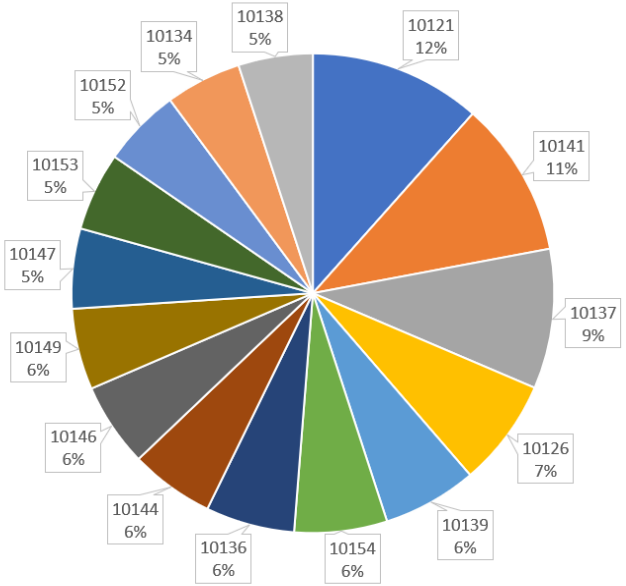
From: Tan, Steinbach, Kumar, *Introduction to Data Mining*, McGraw Hill 2006

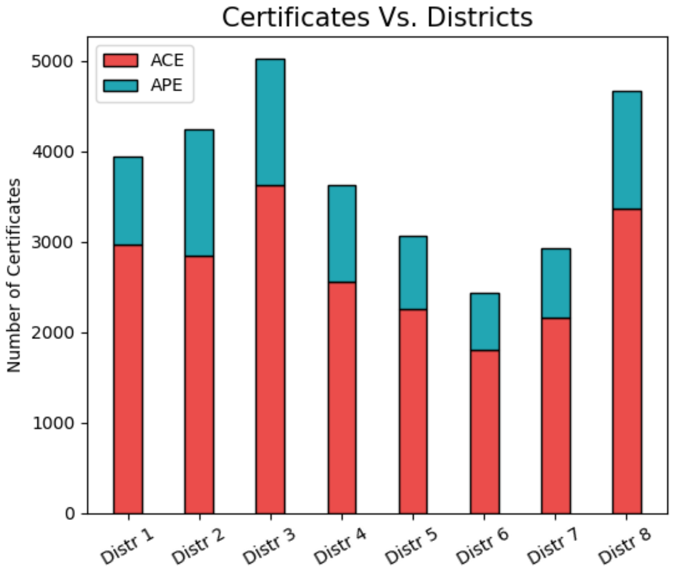


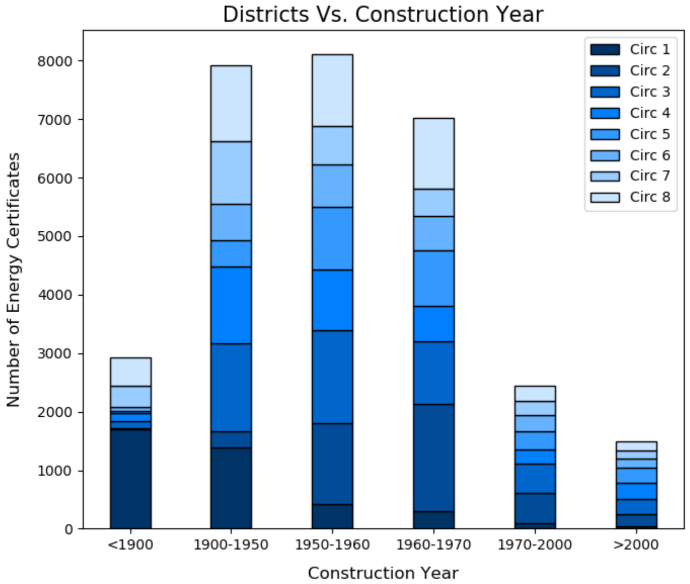
Cleaned dataset related to Turin



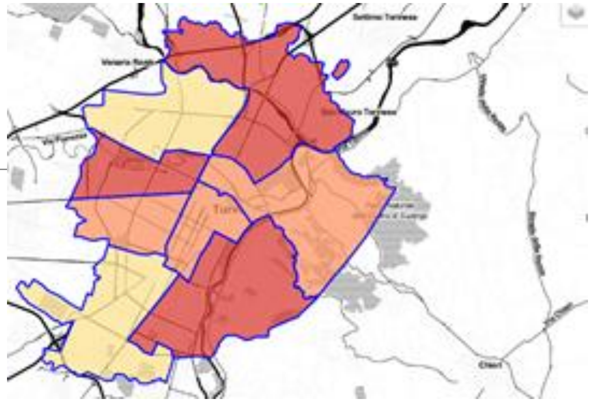
- E1 (1) dwellings in Torino used as **permanent residence**
- EPCs issued in the period: **2009 – 2018**
- EPCs for ***particella, foglio*** *e* ***subalterno*** (identifying each single dwelling)
- **Number of selected EPCs: 29,934**
- **Percentage of EPCs with respect to the total building number in the ISTAT database: 29,934/600,000 ~ 5 %**



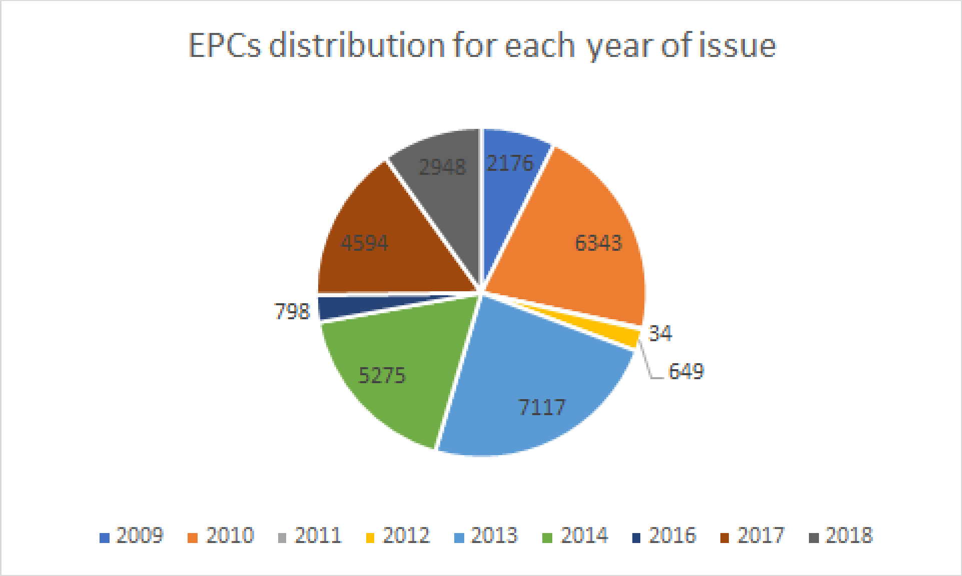


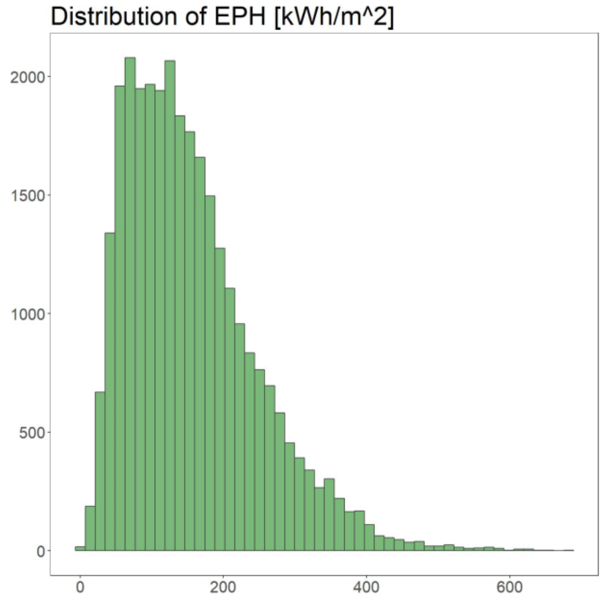


Cleaned dataset related to Turin

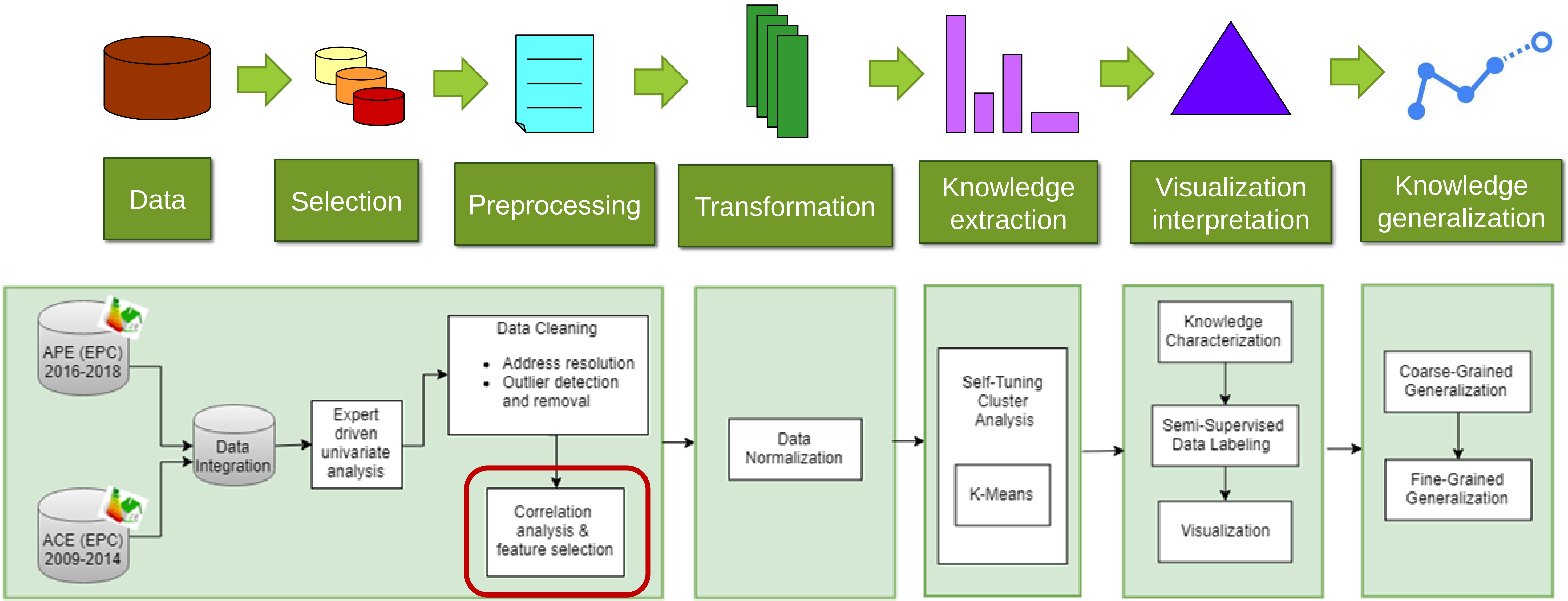


- E1 (1) dwellings in Torino used as **permanent residence**
- EPCs issued in the period: **2009 – 2018**
- EPCs for ***particella, foglio*** *e* ***subalterno*** (identifying each single dwelling)
- **Number of selected EPCs: 29,934**
- **Percentage of EPCs with respect to the total building number in the ISTAT database: 29,934/600,000 ~ 5 %**





Preprocessing-Correlation Analysis

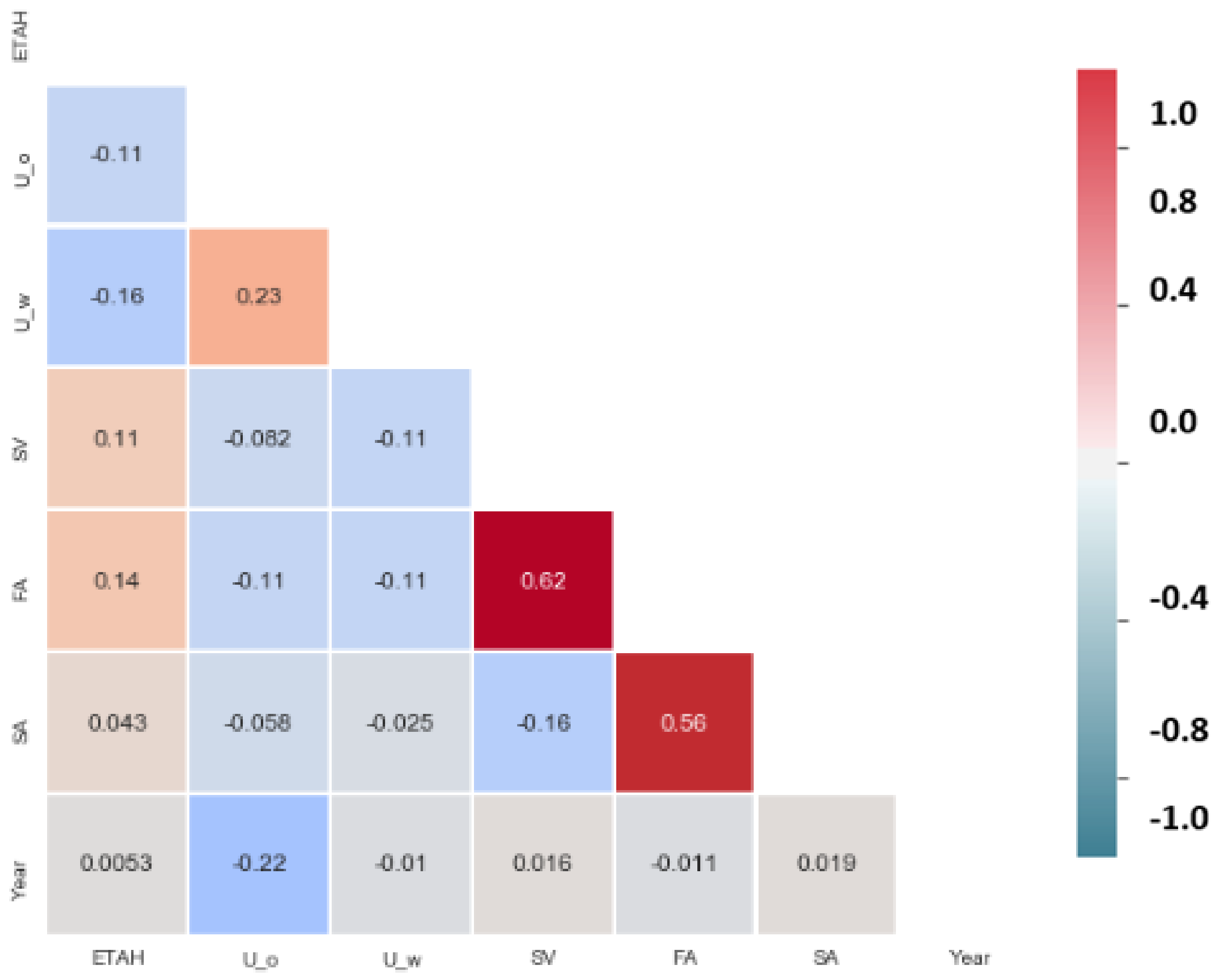


Correlation analysis

Data-driven

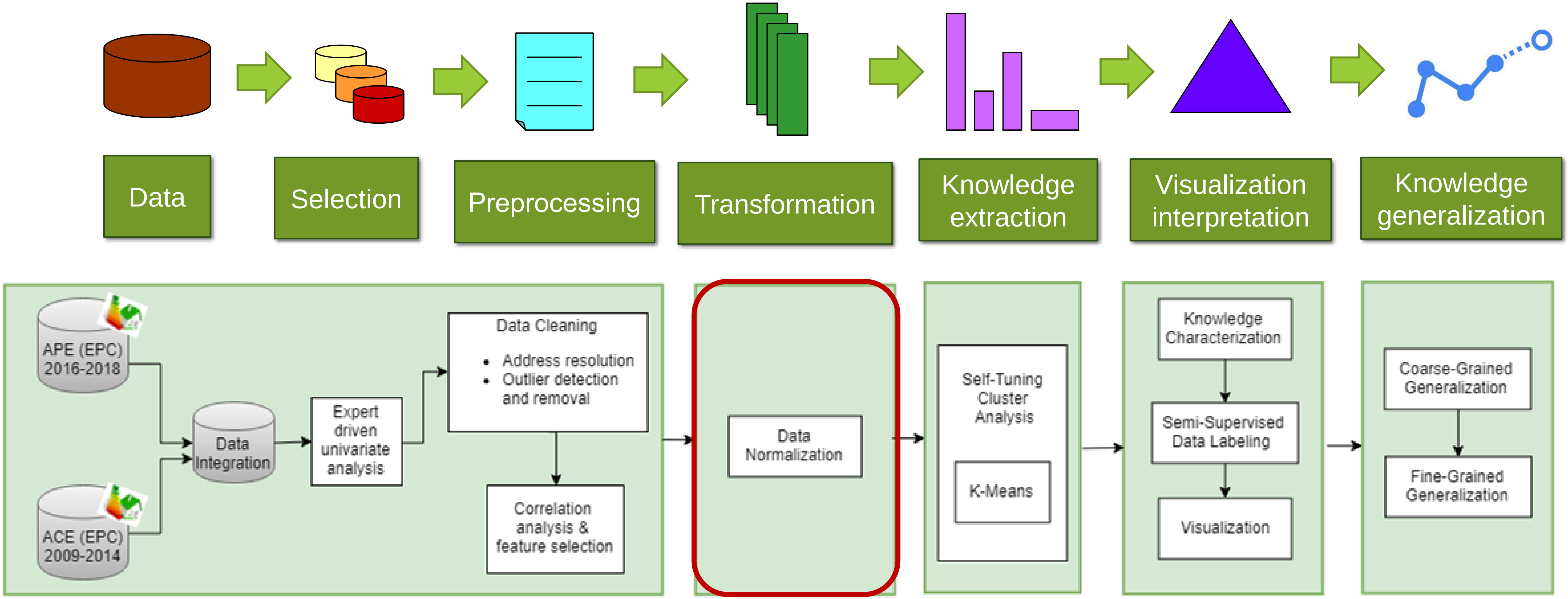
- Feature **removal** (correlation-based approach)
  - simplifying the model computation
  - improving the model performance
- Feature **selection based on correlation test**
  - Features highly-correlated with other attributes could be discarded from the analysis
    - having dependence or association in any statistical relationship, whether causal or not

Correlation analysis

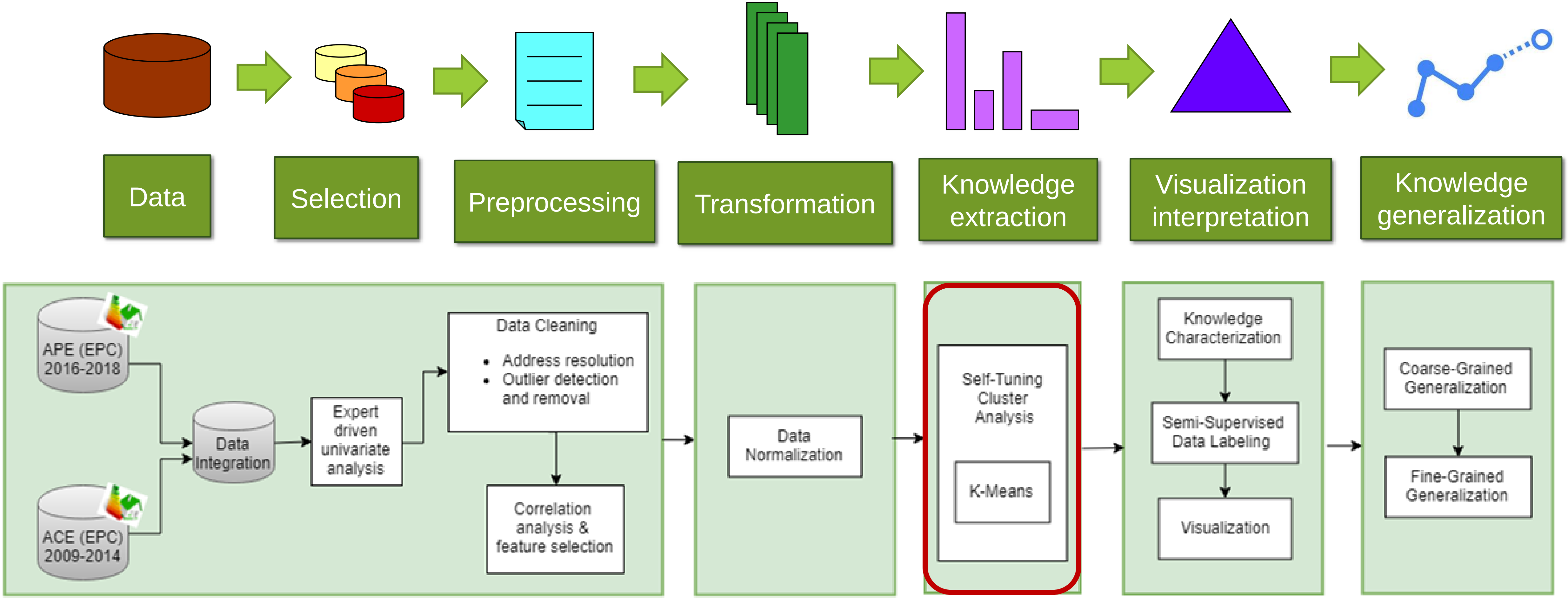


- **S/V** Surface/Volume Ratio
- **U_o** Average U-value of opaque envelope
- **U_w** Average U-value of the windows
- **ETAH** Heating system global efficiency
- **SA** Surface Area
- **FA** Floor Area
- **Year** Construction Year

Transformation



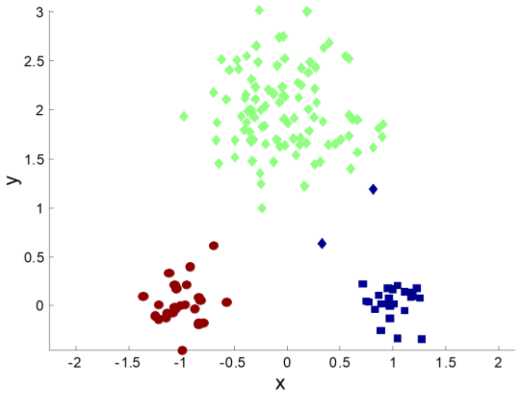
Knowledge extraction process from EPCs



Self-tuning cluster analysis

Clustering algorithms enriched by **self-tuning strategies** (i.e., parameter **autoconfiguration**)

- Partitional algorithm: **K-Means**
  - Each cluster is represented by a **centroid**
  - The desired **number of clusters** is identified by the user



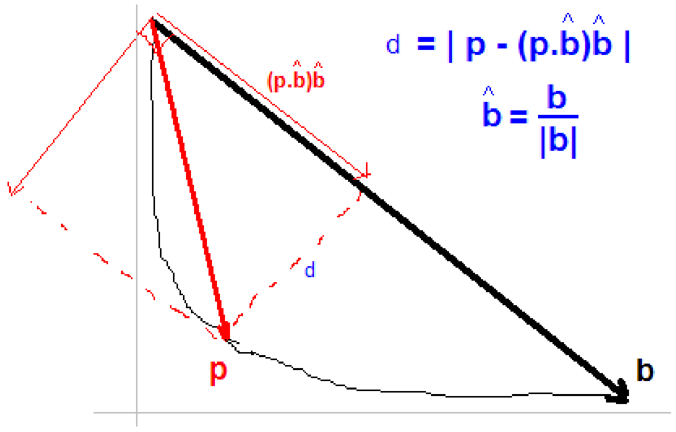
Optimal Clustering with K-Means

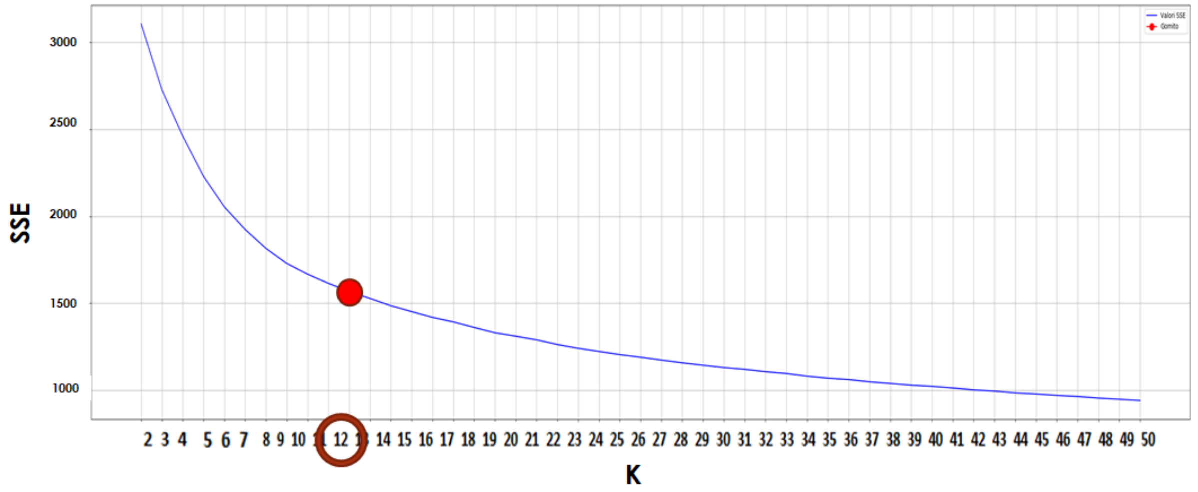
From: Tan, Steinbach, Kumar, *Introduction to Data Mining*, McGraw Hill 2006

Self-tuning cluster analysis:

Clustering algorithms enriched by **self-tuning strategies** (i.e., parameter **autoconfiguration**)

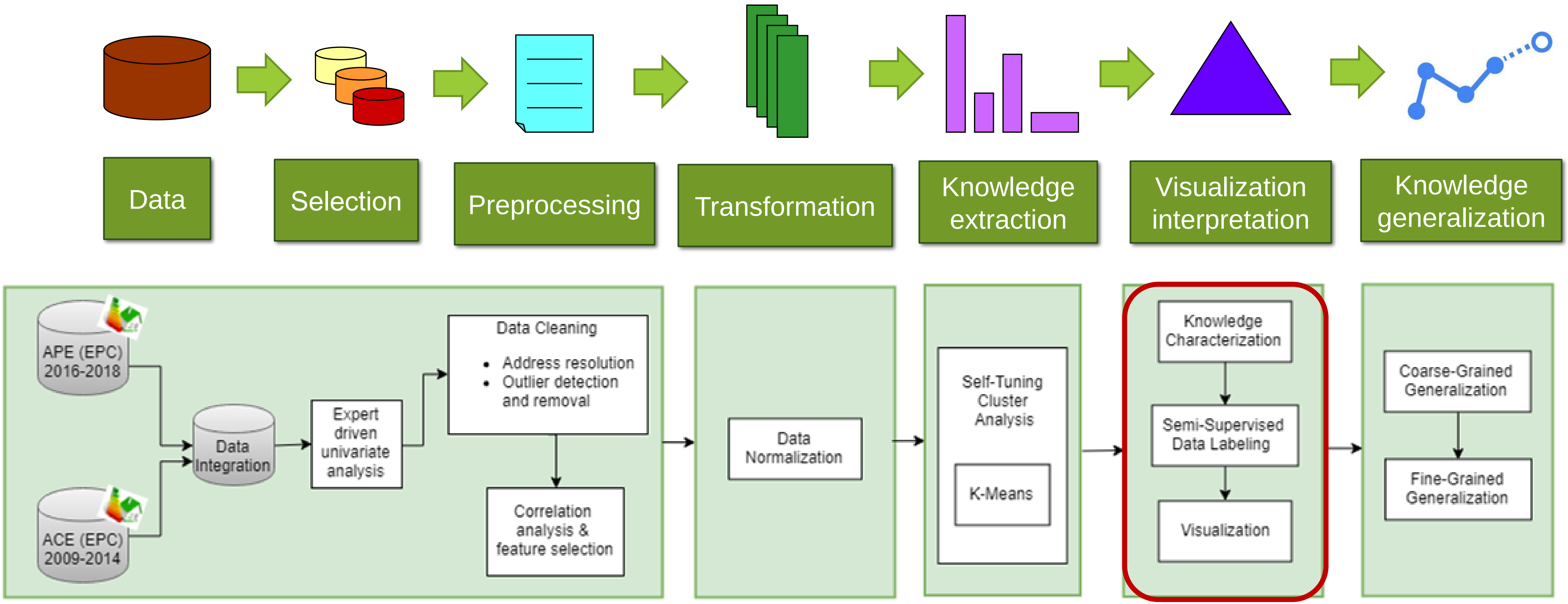
- Partitional algorithm: **K-Means**
  - Each cluster is represented by a **centroid**
  - The desired **number of clusters** is identified by the user
- Self-tuning strategy based on the **Elbow plot:** quality-measure trend (e.g., SSE) vs K
  - *The methodology presented in "Finding a Kneedle in a Haystack: Detecting Knee Points in System Behavior"*, Ville Satopaa; Jeannie Albrecht; David Irwin; Barath Raghavan has been integrated
  - The gain from adding a centroid is negligible
  - The reduction of the quality measure is not interesting anymore





Ville Satopaa et al; *"Finding a Kneedle in a Haystack: Detecting Knee Points in System Behavior"*, 2011 31st International Conference on Distributed Computing Systems.

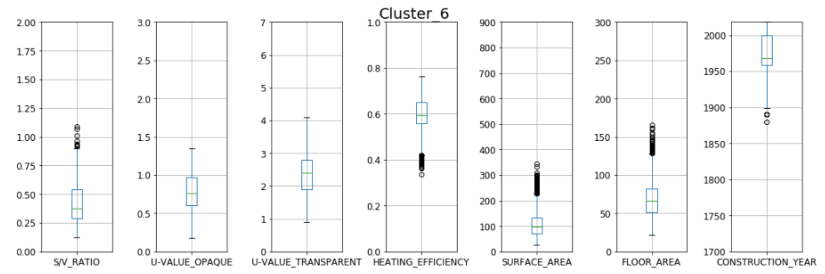
Knowledge extraction process from EPCs

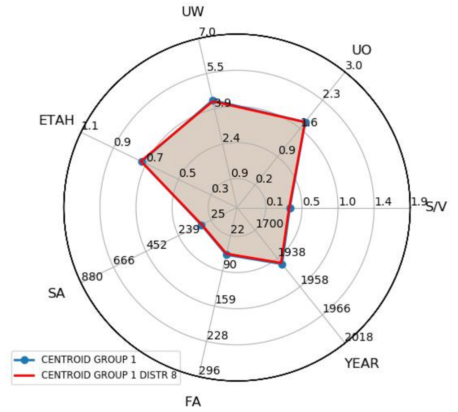


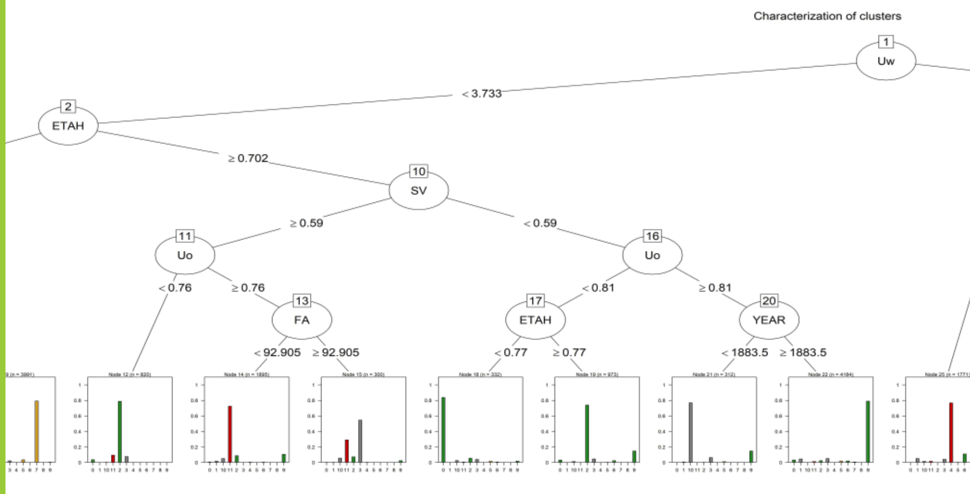
Cluster characterization

Each discovered cluster of EPCs is characterized through:

- Centroids represented through radar plots
- Data distribution for each attribute modeled through boxplot
- Cluster labels, assigned with the support of the domain expert





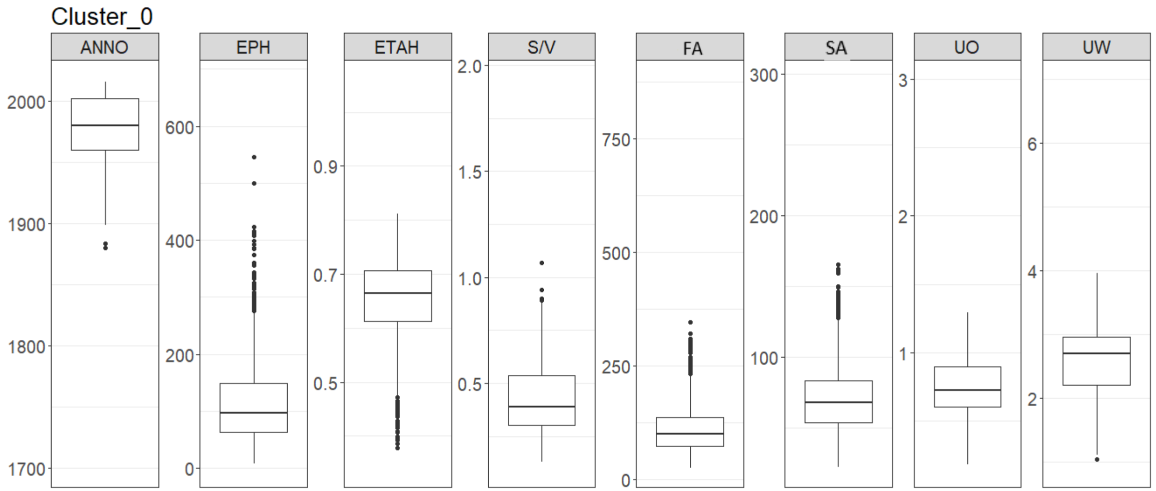


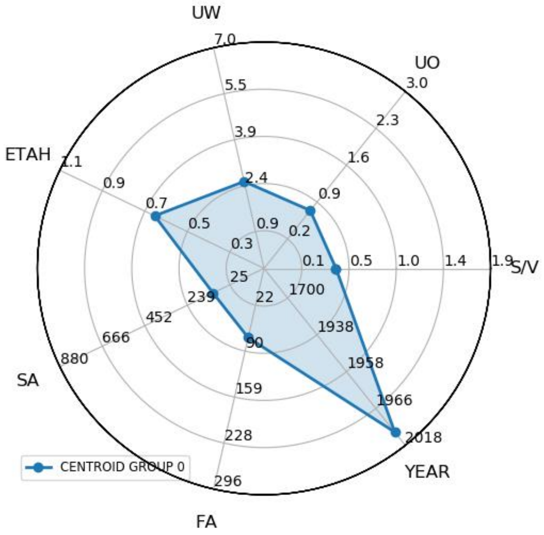
Cluster characterization

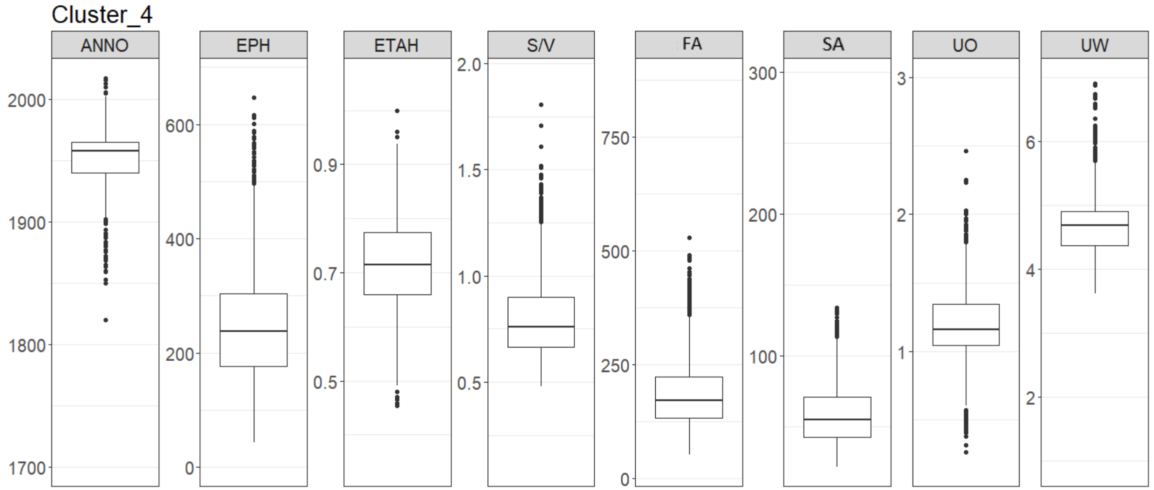
Cluster ID	EPC #
Cluster 0	1,783
Cluster 1	1,810
Cluster 2	1,683
Cluster 3	857
Cluster 4	2,720
Cluster 5	1,450
Cluster 6	4,083
Cluster 7	3,574
Cluster 8	4,916
Cluster 9	3,725
Cluster 10	808
Cluster 11	2,525

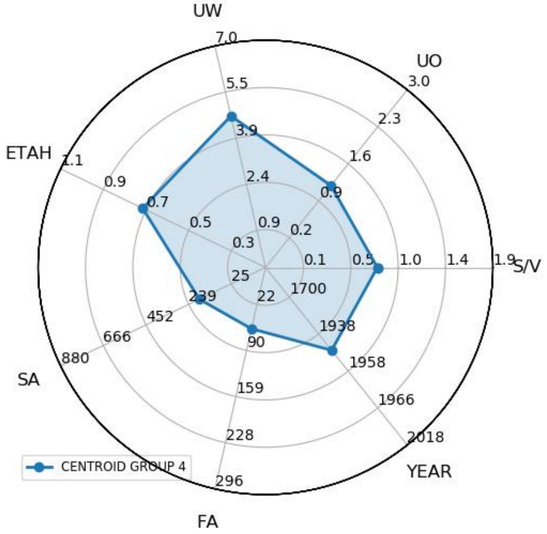
		Districts							
		1	2	3	4	5	6	7	8
Cluster Label	0	101	245	321	217	281	222	172	224
	1	231	289	311	249	131	137	145	317
	2	91	236	264	283	262	111	196	240
	3	251	54	92	79	23	42	109	207
	4	218	395	523	304	306	270	291	413
	5	430	185	234	165	33	37	105	261
	6	383	758	688	472	375	297	360	750
	7	419	433	637	480	415	325	351	514
	8	435	738	860	649	587	450	496	701
	9	480	591	643	472	351	274	359	555
	10	643	2	8	14	1	9	53	78
	11	255	321	440	245	300	268	292	404

Clusters of EPCs: High vs Low energy performance

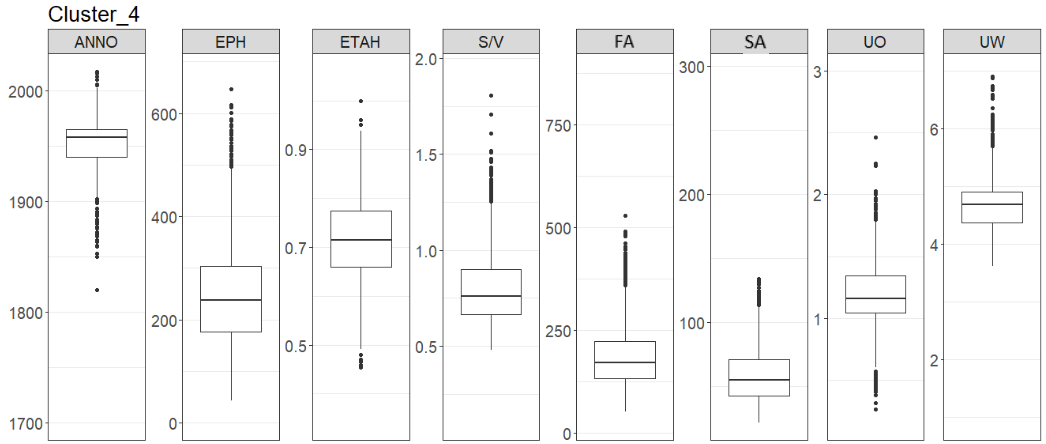


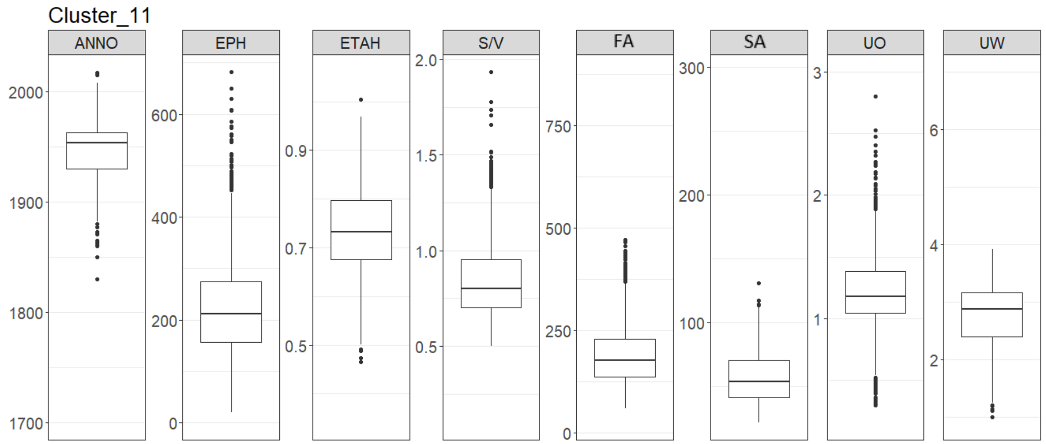


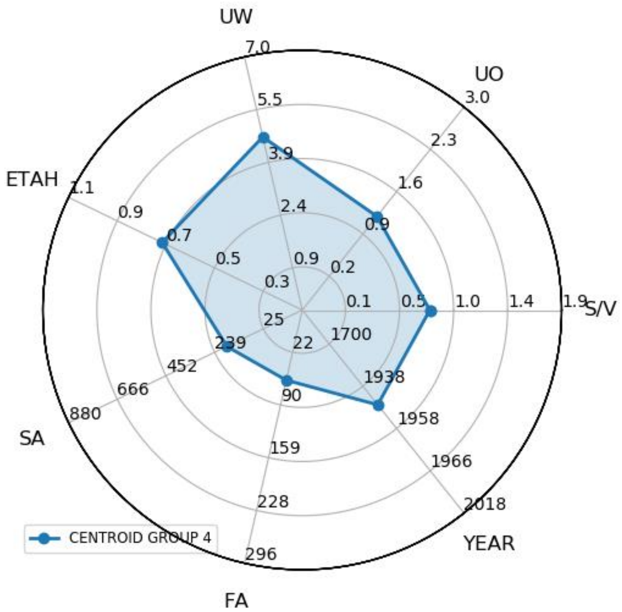


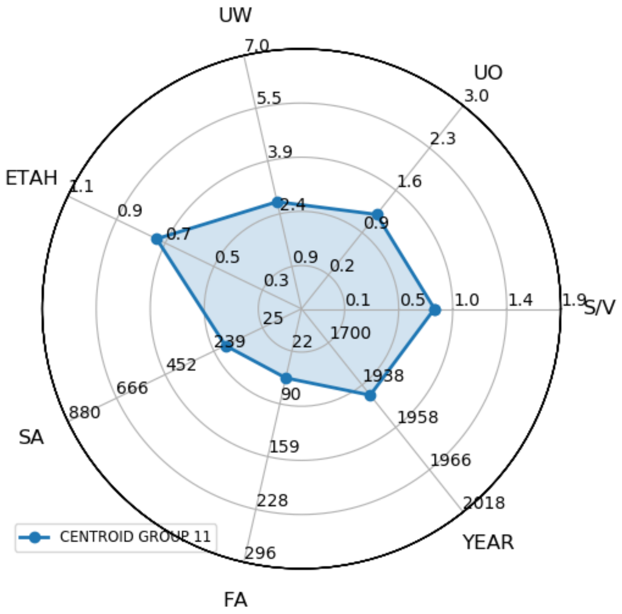


Clusters of EPCs: Low energy performance

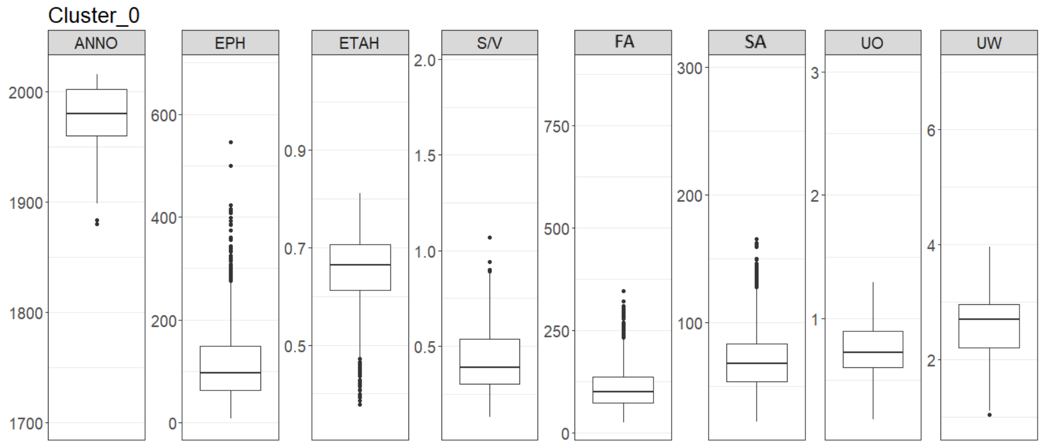


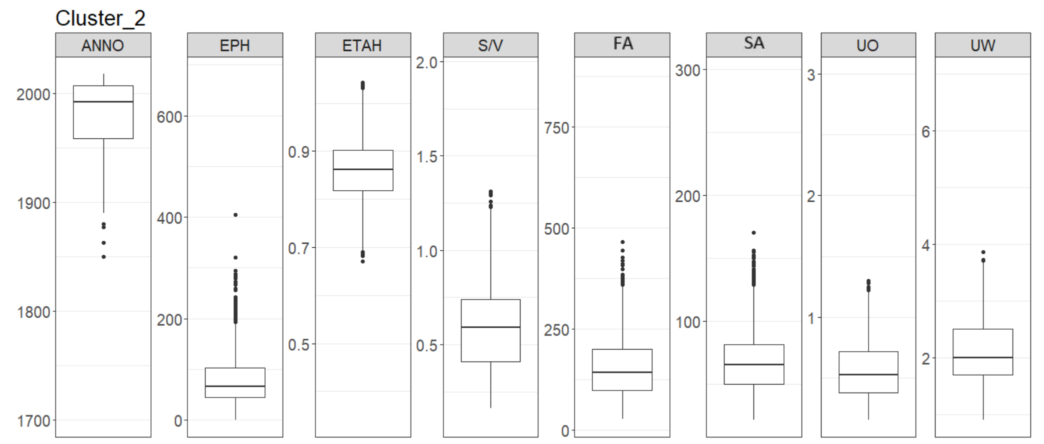


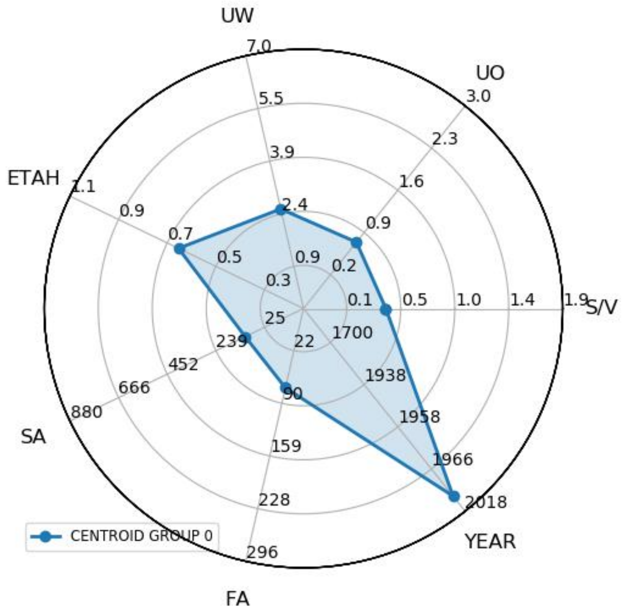


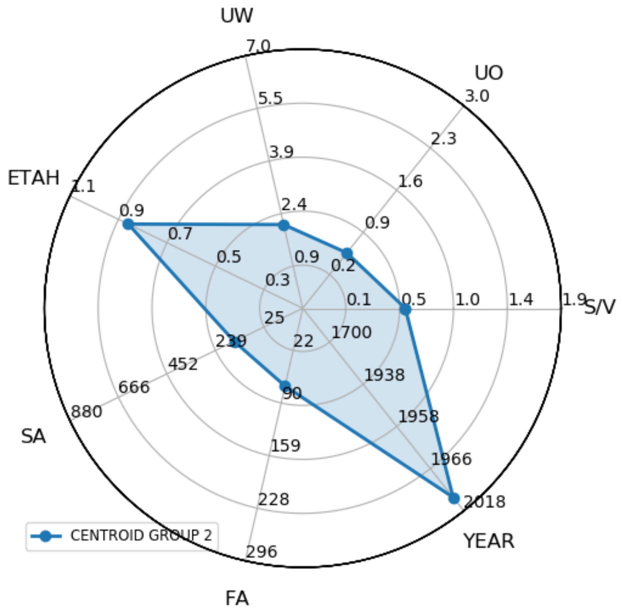


Clusters of EPCs: High energy performance







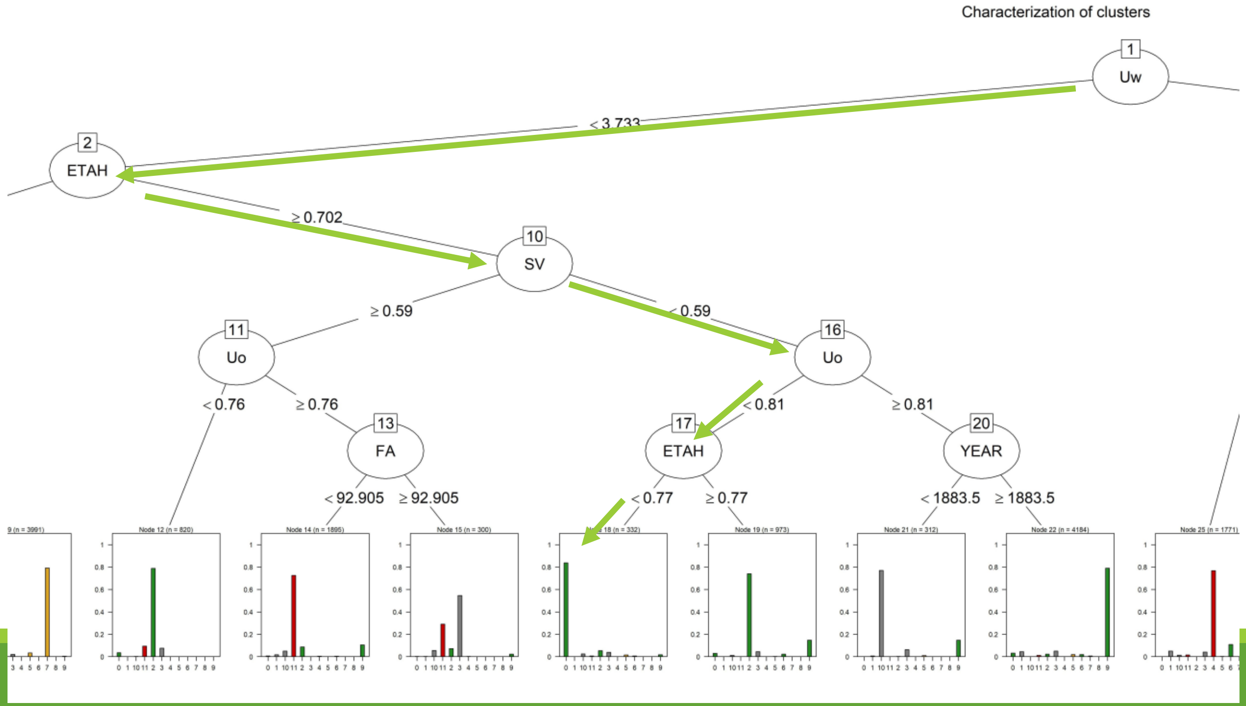


Cluster characterization through CART rules

A CART is built by considering all cluster input variables as input and the cluster id as label to be predicted

- ***Transparent*** *self-describing model, directly "readable" by humans*

Rules are automatically *extracts from CART by visiting its paths, being* ***directly exploitable*** *by all stakeholders (including non-experts) and by the domain expert to define the meaning of each group.*



IF
Uwindow <3.733,
ETAH [0.702-0.77] ,
SV<0.59,
Uopaque<0.81
THEN
ClusterID = 0

Semi-supervised data labeling

ClusterID	Energy Performance Label	Color	Description
0	High		High performing envelope, medium performing energy system
1	X		Low performing envelope, low values of SV
2	High		High performing envelope and energy system
3	X		Buildings with large surface area
4	Low		Low performing envelope, high values of SV
5	Medium		Low performing envelope, medium performing energy system, low values of SV
6	High		Low performing envelope, high performing energy system, low values of SV
7	Medium		High performing envelope, low performing energy system, low values of SV
8	Medium		Medium performing envelope, low performing energy system, low values of SV
9	High		Medium performing envelope, medium performing system, low values of SV
10	X		Historical buildings
11	Low		Medium performing envelope, medium performing system, high values of SV

Knowledge visualization

Maps with different spatial granularity levels

- City
- District
- Neighborhood
- Dwellings

Different types of maps

Choropleth maps

- An aggregation metric is required
  - Majority model
  - Statistical functions to be defined with the domain expert

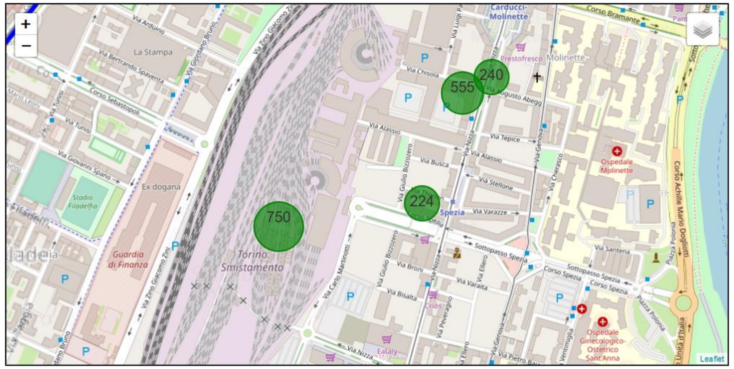
Scatter maps with individual markers

Maps with marker-clusters

- Dynamic plots to model aggregated APEs

TUCANA

Turin Certificates Analysis



Web Application

Knowledge generalization: coarse grained

Two step approach to assign to a new dwelling its cluster label, representing its energy performance:

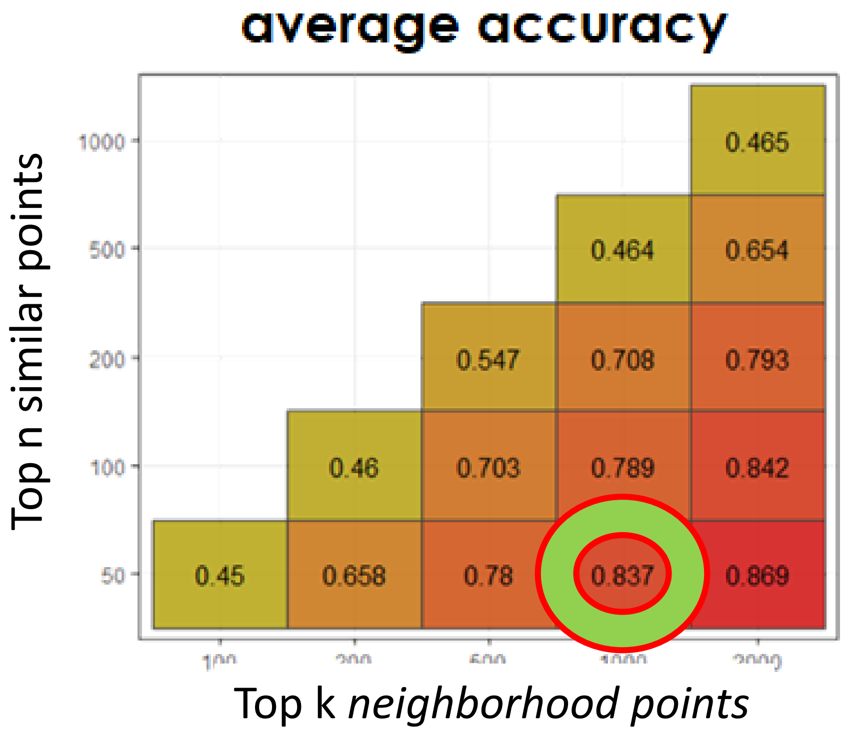
1) Identification of the dwelling neighborhood given a maximum number of dwellings
   A. Given the lat and long of the new dwelling, its closest dwellings are selected
2) K-nearest neighborhood
   A. Among the selected neighbors, the top K similar EPCs [according to the available cluster input variable] are chosen
   B. The cluster label to be predicted is the most frequent label among the ones selected in 2.A

The above methodology can be exploited when:

- All EPC features (considered in the cluster analysis) are available for the new dwelling
- A subset of features (considered in the cluster analysis) is available for the new dwelling
  - Preliminary tests on geometrical dwelling features
- Only latitude and longitude are available for the new dwelling
  - Only steps 1 and 2.B are carried out

Knowledge generalization: coarse grained

All EPC features (considered in the cluster analysis) are available for the new dwellings



*A good **trade-off** is in correspondence of*

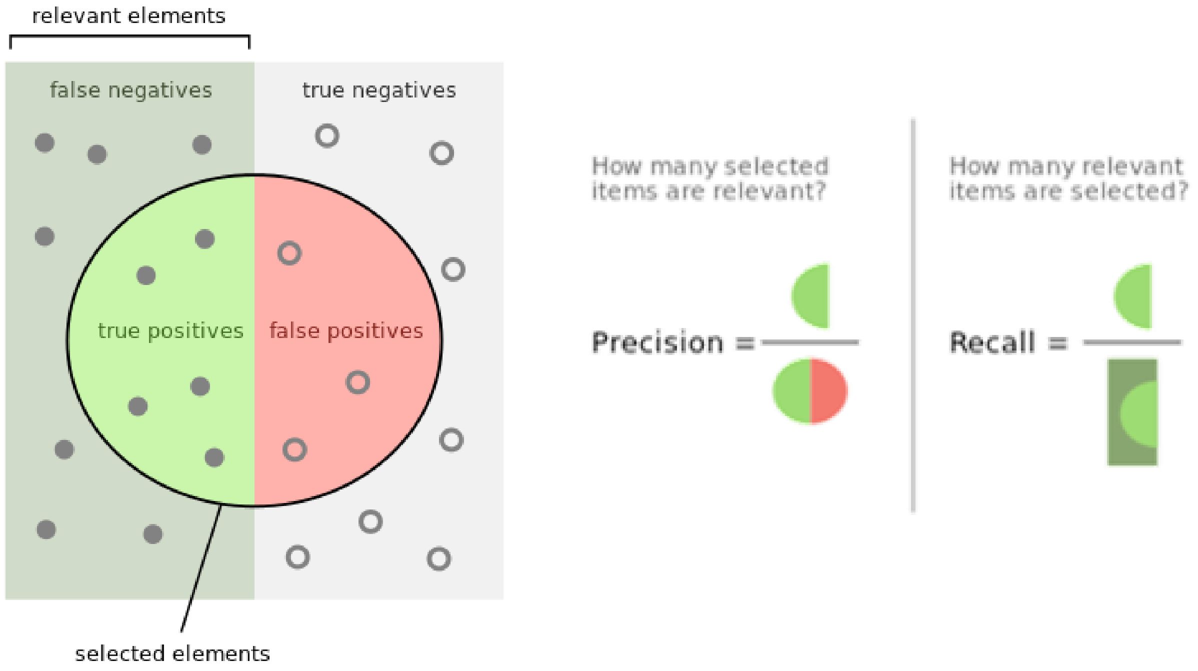
- *number of neighborhood points equals to 1000*
- *number of similar points equals to 50.*

Knowledge generalization: coarse grained

Class	Precision	Recall
0	0.917	0.576
1	0.951	0.480
10	0.963	0.792
11	0.829	0.863
2	0.942	0.662
3	0.962	0.481
4	0.897	0.861
5	0.820	0.580
6	0.839	0.950
7	0.787	0.950
8	0.842	0.989
9	0.765	0.954

Accuracy	Average Precision	Average Recall
0.839	0.876	0.761

*For each cluster, two important **model evaluation metrics** are evaluated.*



Knowledge generalization: coarse grained

Only the geometrical EPC features (considered in the cluster analysis) are available for the new dwelling: SV, Floor Area and Surface Area

Average accuracy

Top n similar points

Top k *neighborhood points*

*A good **trade-off** is in correspondence of*

- *number of neighborhood points equals to 100*
- *number of similar points equals to 25.*

Knowledge generalization: coarse grained

ClusterID	Precision	Recall
0	0.292	0.204
1	0.117	0.051
10	0.381	0.564
11	0.397	0.429
2	0.319	0.261
3	0.790	0.299
4	0.358	0.415
5	0.497	0.467
6	0.245	0.306
7	0.247	0.185
8	0.274	0.413
9	0.203	0.142

Accuracy	Average Precision	Average Recall
0.299	0.343	0.311

Knowledge generalization: fine grained

Predition of the value of one missing cluster input variable

1) A regression model is built on the cleaned dataset by analyzing a subset of cluster input variables
2) Different algorithms were integrated:
    1) LASSO
    2) RIDGE
    3) K-NN regressor
    4) Polinomyal regressor
    5) Support Vector regression
3) 10-fold cross validation has been exploited to compute the quality metrics and select the best algorithm

The above methodology can be exploited before applying the coarse-grained generalization approach

Cluster input variables are characterized by a low value of correlations

- **Strong point** to obtain good quality model by means of the cluster analysis
- **Weak point** to build an accurate regression model able to predict one of the cluster input variable based on the others

Knowledge generalization: fine grained

Experiment ID	Input Variables	Predicted Variable	Regression model	Quality metric R^2
1	ETA_D, ETA_G, ETA_R, U_o, U_w, FA, SA, Year, SV	ETAH*	Lasso regressor	0.97
2	ETA_E, ETA_G, ETA_R, U_o, U_w, FA, SA, Year, SV	ETAH*	Lasso regressor	0.91
3	U_o, U_w, FA, SA, Year, ETAH	SV	K-NN regressor	0.85

*ETAH. This index considers the efficiency of each subsystem of the dwelling: generation subsystem (ETA_G), distribution subsystem (ETA_D), emission subsystem (ETA_E) and control subsystem (ETA_R)

Experiment ID	# EPCs
1	317
2	405
3	87

Joint publications

Cerquitelli T., Di Corso E., Proto S, Capozzoli A., Bellotti F., Cassese M.G., Baralis E., Mellia M., Casagrande S., Tamburini M., *Exploring Energy Performance Certificates through Visualization*. In Proceedings of the Workshops of the EDBT/ICDT 2019 Joint Conference (EDBT/ICDT 2019) Lisbon, Portugal, March 26, 2019.

Cerquitelli T., Di Corso E., Proto S, Capozzoli A., Mazzarelli D. M., Nasso A., Baralis E., Mellia M., Casagrande S., Tamburini M., *Visualising high-resolution energy maps through the exploratory analysis of energy performance certificates.* Accepted for publication, to be presented at SEST 2019, Porto, Portugal, September 9-11, 2019.

Public talks

Tania Cerquitelli ***Creare valore e strutturare conoscenza a partire da open data energetici: metodi, sfide e opportunità.*** Open Access Week @ POLITO, October 23th, 2018 Turin, Italy http://www.politocomunica.polito.it/news/allegato/(idnews)/11788/(ord)/0

Tania Cerquitelli ***Visualizing high-resolution exploratory energy maps by analyzing energy-performance certificates*** The 4th Workshop of the SmartData@PoliTO Interdepartmental Center will be held on February 28th, 2019 at Politecnico di Torino – AULA MAGNA https://smartdata.polito.it/4th-smartdata-workshop-public/#cerquitelli

Tania Cerquitelli and Alfonso Capozzoli Exploring open data to spread out knowledge: a real-world use case in the energy domai. Focus on Open Access, Università di Torino, May 7th, 2019 Turin, Italy.
http://www.politocomunica.polito.it/en/news/allegato/(idnews)/12677/(ord)/0

… questions?

Tania CERQUITELLI