IMPROVING ENERGY MODELING OF LARGE BUILDING STOCK THROUGH THE DEVELOPMENT OF ARCHETYPE BUILDINGS

Ilaria Ballarini, Stefano Paolo Corgnati, Vincenzo Corrado, and Novella Talà
TEBE Research Group, Department of Energetics, Politecnico di Torino, Torino, Italy
ilaria.ballarini@polito.it - stefano.corgnati@polito.it - vincenzo.corrado@polito.it – novella.tala@polito.it

ABSTRACT

In this paper a selection process, based on statistical techniques, of representative buildings is presented. Starting from a real estate stock it is possible to draw a sample and calculate the relevant sample statistics. As second step, their elaboration permits to pick out real buildings with geometrical and thermo-physical characteristics similar to the average of the building sample. In addition, the results of this method are compared to those obtained by segmentation (or cluster) analysis, that is a method to partition a set of houses into groups having similar profiles.

Finally, using the Piedmont Regional Database of Energy Performance Certificates these approaches are applied in order to verify the reliability of the analyses proposed. Potentialities and limitations of the performed analyses are critically discussed, as well.

INTRODUCTION

The reduction of energy consumption and the associated greenhouse gas emissions in every sector of the economy is a very topical research item. The residential sector consumes a large amount of energy in every country and therefore it deserves particular attention.

Comprehensive energy models are needed to assess the effects of new energy efficient technologies on residential housing systems and to identify potential improvements on subsystems.

The first step to develop a large building stock energy model is to define reference buildings that represent certain categories within stock identified according to predetermined criteria and reflect the entire existing stock.

Indeed, to this aim, it is fundamental the application of a methodology for the definition of "building types", which allows the classification of existing buildings in categories ("buildings-types") to be analyzed and investigated.

Starting from the energy demand (calculated or measured) of the typical building and highlighting its representativeness within the stock, it is possible to estimate the entire stock consumption. Thus the energy requirements of the stock is re-estimated using retrofitted building types and the potential savings achievable through retrofit actions addressed to building envelope and heating systems are obtained by simple difference between ante and post retrofit scenarios.

This study is a part of TABULA (*Typology Approach for Building stock Energy Assessment*) project within the European program "Intelligent Energy Europe" (IEE).

The TABULA project objective is to create a concerted structure on the building typologies in Europe in order to estimate the energy demand of residential building stocks at national level and, consequently, to predict the potential impact of energy efficiency measures and to select effective strategies for upgrading existing buildings.

In order to define a typical house useful for describing the thermal and geometric characteristics of a group of houses, the first step consists of identifying independent variables influencing the multitude of parameters that are specific to the building.

The TABULA project has fixed three independent variables which are: location, age and geometry (shape/volume). In the specific Italian case, the three-dimensional space that generate appropriate reference building includes 3 climatic zones, 8 ages and 4 geometries of Italian housing (single-family house, multi-family house, terraced house, apartment block), the combinatorial process produces 96 building typologies (Parekh, 2005).

Three approaches have been proposed to define building typologies.

According to the first approach the representative building (Example Building) relevant to construction period, geometry and region is defined according to an expert choice based on rules-of-thumb to compensate the lack of statistical information. The second method identifies the typical building (Real Building) elaborating statistically data to extrapolate the real building with geometrical and thermo-physical characteristics similar to the average of the building sample. The third method identifies the typical building (Theoretical Building) as a building that is the most probable of a group of buildings.

SOURCE OF DATA

Building energy performance (EP) certification in Piedmont

The database contains records for more than 66.000 houses rated across Piedmont.
The 66.000 house records represent the result of the information collected by EP certification schemes.
The database contains information on physical characteristics and calculated energy requirements of each house. Each submission includes more than 40 information fields. The data includes:

- location;
- construction period;
- form;
- heated gross volume;
- net floor area;
- window average thermal transmittance;
- calculated energy demands and indicators.

The purpose of the EPCs database is also to gather the individual energy analyses data. Once an energy advisor successfully completes the energy assessment of a house, the resulting energy analysis data is collected and stored into the database (Blais *et al.*,2005).
The energy performance index, based on asset rating approach, is evaluated by means of software tools based on Italian technical specifications UNI/TS 11300: it represents the estimated energy demand to meet the different needs associated with a standardized use of the building. No actual energy consumptions are reported in the EP certification: this do not represent a limitation for this study because the availability of energy performance under standard conditions in suitable for the definition of "reference buildings". On the contrary, the knowledge of actual energy consumptions is fundamental to calibrate energy calculations when specific energy savings actions are investigated (e.g. related to actual energy functioning as building operation set-points, occupant behaviors, etc.)

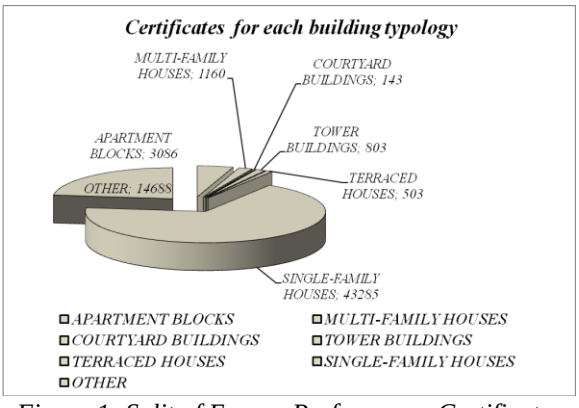

Figure 1: Split of Energy Performance Certificates for each building typology (66063 certificates).

In order to validate the quality of data and to simplify the analysis the amount of data is restricted to only 7104 certificate schemes. In particular, apartment blocks, multi-family houses, terraced houses and single-family houses have been considered. Such data are conveniently illustrated by means of the pie charts in Figure 1 and Figure 2.
The application of the third method for generating reference buildings is still ongoing. It finally aims to develop building types libraries providing appropriate default values for detailed energy analyses (Sansregret *et al.*, 2009).

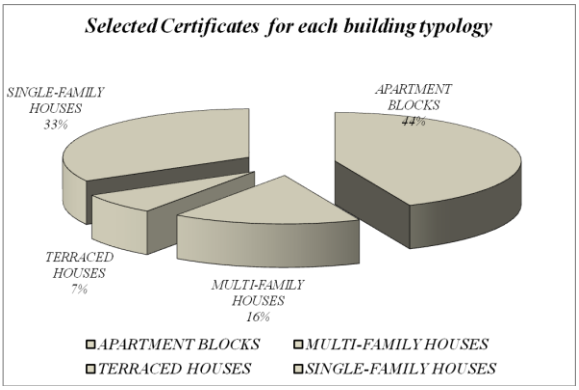

Figure 2: Split of the selected Energy Performance Certificates for each building typology (7104 certificates).

Interpretation of data and removal of outliers

The descriptive statistics are a way to summarize such data into a few numbers that contain most of the relevant information.
The measures of location permit to locate the data values on the number line.
Data entry errors also called outliers are anomalous and incoherent observations respect to the other observations in a data set. The outliers exist in almost all real data and the sample average is sensitive to these problems. One bad datum can move the average away from the center of the rest of the data by an arbitrarily large distance while the median is a measure that is robust to outliers. The median is the 50^{th} percentile of the sample, which will only shift slightly if a large perturbation occurs.
The purpose of measures of spread is to understand how the data values are dispersed on the number line. The difference between the maximum and minimum values is the simplest measure of spread. But the range is not robust to outliers.
The standard deviation and the variance are common measures of spread that are optimal for normally distributed samples.
The Interquartile Range (IQR) is the difference between the 75th and 25^{th} percentile of the data. Since only the middle of the data affects this measure, it is robust to outliers.
The data contained into the database have been summarized statistically.
In this context sample data may be represented by a variety of different types of distribution, it is most

appropriate to evaluate the median and mid-quartile ranges (Mortimer *et al.*, 1999).
In particular, frequency distributions of energy needs for heating can be plotted and statistical analysis can be used to quantify the trend of characteristics for samples of given buildings. Figure 3 shows the frequency distribution of energy needs for heating for the whole sample of terraced houses. Figure 4 shows the box-plot of energy needs for heating of Piedmont housing. As shown, there have been slight decreasing over the last three periods.
The median indicates the point in the frequency distribution which equally divides the total number of data points; 50% of the data occur on either side of the median. The mid-quartile range covers the middle 50% portion of the sample data. The medians and mid-quartile ranges of the energy needs for heating (Q_H) for terraced houses from a sample of 325 are presented in Table 1.

Table 1
Means, Medians, 25th and 75th percentiles of the energy needs for heating (Q_H) for terraced houses from a sample of 325in the Piedmont.

Construction period	Mean	Q_1	Median	Q_3
I (<1900)	268	193	242	319
II (1901-1920)	194	111	194	271
III (1921-1945)	240	161	210	263
IV (1946-1960)	231	156	237	307
V (1961-1975)	434	161	207	383
VI (1976-1990)	158	113	161	194
VII (1991-2005)	116	99	112	127
VIII (>2005)	74	51	64	94

In particular, this dataset is partitioned into eight building age classes containing the following numbers of houses (Table 2):

Table 2
Distribution of the sample in age classes.

Class	No. of houses	Period
I	89	before 1900
II	31	1901-1920
III	52	1921-1945
IV	29	1946-1960
V	35	1961-1975
VI	24	1976-1990
VII	35	1991-2005
VIII	30	after 2005

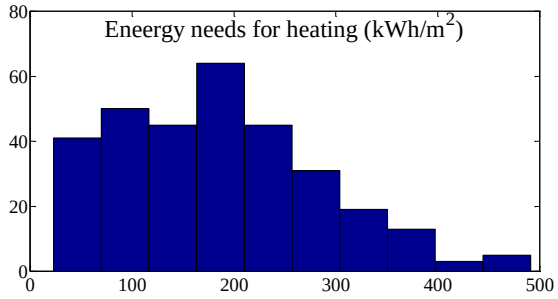

Figure 3: Energy needs for heating for the sample of terraced houses.

As shown in Table 1 (also confirmed by the box-plot in Figure 4), the mean and median are not so close when outliers exist.
As above, outliers are atypical and inconsistent observations with regard to majority of observations in a data set. In statistics there are several means to pick out outliers, for example, with the z-scores and the box and whisker plot.
In the representation of the box and whisker plot, outliers are reported individually being an anomaly respected to distribution data.
Likely cases of outliers in a dataset are:

- errors due to measurements, generated by differences between the sensitivity of instruments and researchers' skills;
- several environmental factors (time, climate, location, ...) affecting the observations.

The method of identifying invalid data and outliers is described in literature (Tukey, 1977).
Let us denote with *IQR* the interquartile range

$$IQR = Q_3 - Q_1 \quad (1)$$

where:
Q_1 is the 25th percentile and Q_3 is the 75th percentile.
Observations Y that satisfy the following conditions are removed:

$$Y < LAV; \quad Y > UAV \quad (2)$$

with:

$$LAV = Q_1 - 1{,}5 \cdot IQR \quad (3)$$

$$UAV = Q_3 + 1{,}5 \cdot IQR \quad (4)$$

where *LAV* and *UAV* denote the lower adjacent value and the upper adjacent value, respectively.
Prior to any evaluation of the sample, any possible errors detected in the data set have been set aside.

Applications to Energy Performance Certificates

The Piedmont Regional Database of Energy Performance Certificates (EPCs) has been used to define the building typologies within the categories single family houses and terraced houses.

Based on the available data, representative parameters of geometric and thermal features have been selected. These parameters are:

- volume;
- net floor area;
- envelope area to volume ratio;
- number of storeys;
- number of apartments;
- opaque envelope average thermal transmittance;
- window average thermal transmittance.

For each parameter the number of data were sufficient (at least 10 observations) to calculate statistical functions such as mean, median, 25th percentile, 75th percentile. Interquartile ranges (IQRs) are evaluated for all the parameters. The intersection of all IQRs permits to select the single real building whose parameters are the closest to the median values.

$$\left\lfloor IQR_{n_L} \cap IQR_{n_{DW}} \ldots IQR_{U_{OP}} \cap IQR_{U_W} \right\rfloor_{CP_i, BS_j, L_k} \quad (5)$$

where, referring to the Italian part of the TABULA project, the following variables are:

- CP_i denotes the construction periods for i=1,8
- BS_j is the building sizes for j=1,4
- L_k represents the locations for k=1,3

If this procedure gives more than one or no real building, IQRs can be tuned by means of suitable criteria in order to pick out only one real building.

Table 3
Real terraced houses identified by means of the intersection of all IQRs.

Construction period	Q_H **[kWh/m^2]**
I:<1900	237,4
II:1901-1920	193,9
III:1921-1945	183,1
IV:1946-1960	103,9
V:1961-1975	246,3
VI:1976-1990	196,1
VII:1990-2005	118,5
VIII:>2005	48,9

Cluster analysis

Cluster analysis is a technique to partition a set of objects into clusters, in such a way that the profiles of objects in the same group are very similar and the profiles of objects in different groups are quite distinct (Gaitani *et al.,* 2010).

Cluster analysis is performed on each building age class of terraced houses defined in the previous section.The following variables are used:

- energy needs for space heating (Q_H);
- primary energy for space heating (EP_i);
- net floor area (A);
- opaque envelope average thermal transmittance (U_{op});
- window average thermal transmittance (U_w).

The values in the data set are normalised before calculating the distance information because variables are measured against different scales.

These discrepancies can distort the proximity calculations.

The process of normalisation has been performed using the zscore function implemented in MATLAB (Jones, 1996) that converts all the values in the data set to use:

$$Y_z = \left(\frac{Y - mean(Y)}{std(Y)} \right) \quad (6)$$

Figure 5 shows the frequency distributions of the variables considered using the data set of building age VII. In addition, outliers hightlighted in the boxplot (see Figure 6) are removed by the application of the condition in equation (2).

Once the outliers are removed and the data are normalized, the cluster analysis can be carried out. Such analysis is based on the calculation of the distance between every pair of objects in the data set. There exist five metrics to calculate the distance. The result of this computation is commonly known as a similarity matrix (or dissimilarity matrix).

Once the proximity between objects in the data set has been computed, the objects in the data set are separated into clusters. Using several algorithms available in MATLAB, pairs of objects that are close are linked together into binary clusters (clusters made up of two objects) then these newly formed clusters are linked to other objects to create bigger clusters until all the objects in the original data set are linked together in a hierarchical tree.

The hierarchical cluster tree is most easily understood when viewed graphically. The dendrogram represents this hierarchical tree information as a graph, as in the Figures 7 and 8.

In particular, Figure 7 presents the data of the building age class VII organized in 30 clusters . A reduced number of cluster is shown in Figure 8. Such clusters correspond to the intersection of the dendrogram in Figure 7 with an orizontal line such that only five intersections take place.

This procedure permits to identify the cluster containing the representative house for the entire building age class. The reference building of the class is choosen as median value with regard the Q_H.

Among the clusters identified, we select the larger because it contains terraced houses having similar features.

One way to measure the validity of the cluster information is to compare the cophenet correlation coefficent. The cophenet function (implemented in MATLAB) compares two sets of values and computes their correlation, returning a value called the cophenetic correlation coefficient.

The cophenetic correlation coefficient has been used to compare the results of clustering using the data set of building age VII with different distance calculation methods and linkage algorithms (see Table 4).

As shown in Table 4, the higher cophenetic correlation coefficient is obtained using the distance information 'Mahalanobis' in conjuction with the linkage algorithm 'Centroid'. This method has been used to perform the cluster analisys. Results are collected in Table 5.

Table 4
Analysis of cophenetic correlation coefficient.

Metrics	**Linkage algorithms**	**Cophenetic correlation coefficient**
Euclid	Single	0,2171
Standardised Euclid	Complete	0,2171
City block	Average	0,2294
Mahalanobis	Centroid	0,3566
Mikowski	Ward	0,2171

Table 5
Terraced houses identified by means of cluster analysis.

Real terraced houses	Q_H **[kWh/m^2]**
I:<1900	239,7
II:1901-1920	190,4
III:1921-1945	187,6
IV:1946-1960	99,3
V:1961-1975	250,9
VI:1976-1990	187,3
VII:1990-2005	110,0
VIII:>2005	49,2

The comparison of table 3 and 5 allows us to show that the two methods produce similar results in terms of Q_H.

CONCLUSION

The main goals of this paper include the definition of a method to choose the representative buildings as well as an energy analysis, which can be used to classify the energy performance for space heating.

A statistical analysis has been performed on a data set of a total of 325 terraced houses, which represents 7% of the residential houses extracted from the EPCs database in Piedmont. The database contains information on each house with regard to its physical characteristics and energy use.

The classification method here proposed is based on the application of hierarchical clustering techniques.

Results provided by this method are compared to those obtained throught statistical analysis.

This comparison allows to show that the two methods produce similar outcomes in terms of energy need for space heating.

ACKNOWLEDGEMENT

This work was carried out within TABULA (*Typology Approach for Building stock Energy Assessment*) project, financed by the European Commission within the European program "Intelligent Energy Europe".

REFERENCES

Blais S., Parekh A., Roux L. Energuide for houses database-An innovative approach to track residential energy evaluations and measure benefits. Proceedings of the 9th International IBPSA Conference, August 2005.

Gaitani N., Lehmann C., Santamouris M., Mihalakakou G., Patargias P.. Using principal component and cluster analysis in the heating evaluation of the school building sector. Applied Energy 87 (2010) 2079–2086.

Intelligent Energy Europe (IEE). Typology Approach for Building Stock Energy Assessment (TABULA), in Description of the Action, Annex I, SI2.528393, April 2009.

Jones B. MATLAB statistics tool book. The Math Works; 1996.

Mortimer N.D., Ashley A., Elsayed M., Kelly M.D., Rix J.H.R.. Developing a database of energy use in the UK non-domestic building stock. Energy Policy 27 (1999) 451-468.

Parekh A., Development of archetypes of building characteristics libraries for simplified energy use evaluation of houses. Proceedings of the 9th International IBPSA Conference, August 2005.

Sansregret S., Millette J. Development of a functionality generating simulations of commercial and institutional buildings having representative characteristics of a real estate stock in Québec (Canada). Proceedings of the 11th International IBPSA Conference, July 2009.

Tukey J W. Exploratory Data Analysis. Addison-Wesley, 1977.

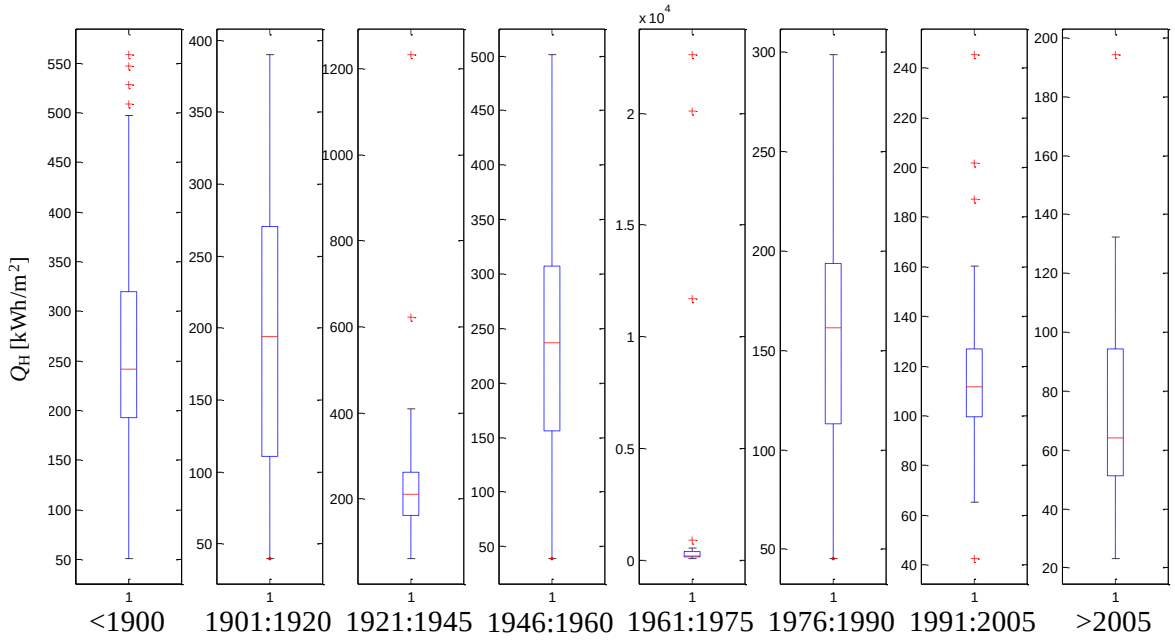

Figure 4: Energy needs for heating for the Piedmont housing. Bottom of the bar is at 25th and top is at 75thpercentiles).

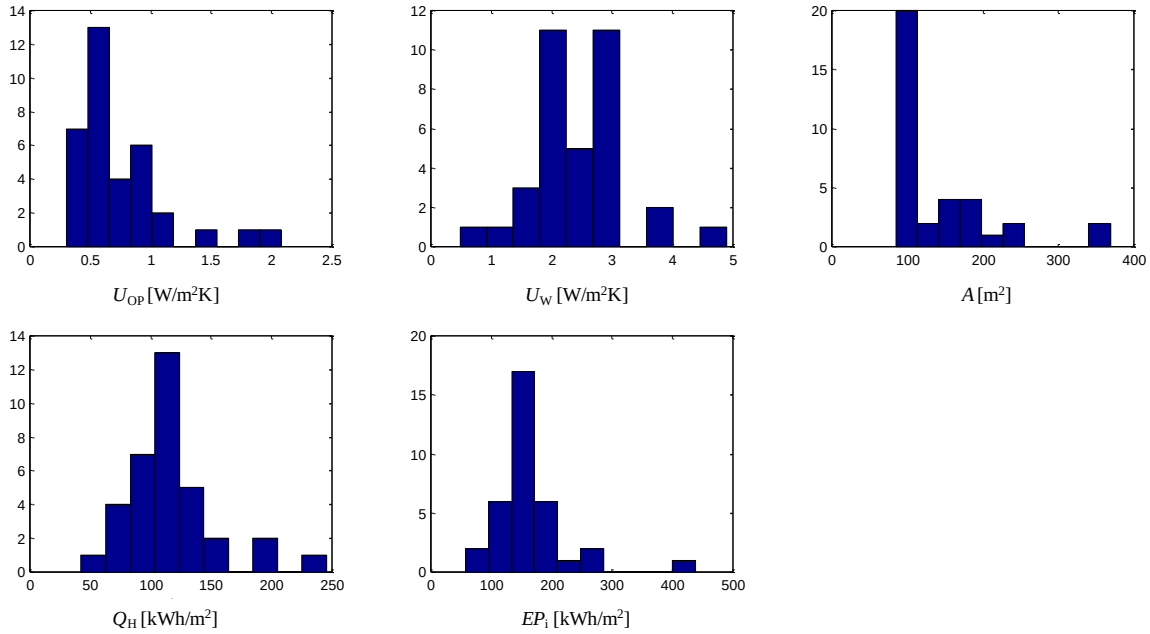

Figure 5: Frequency distributions of the variables considered (U_{op}, U_w, A, Q_H ,EP_i) using the data set of building age VII.

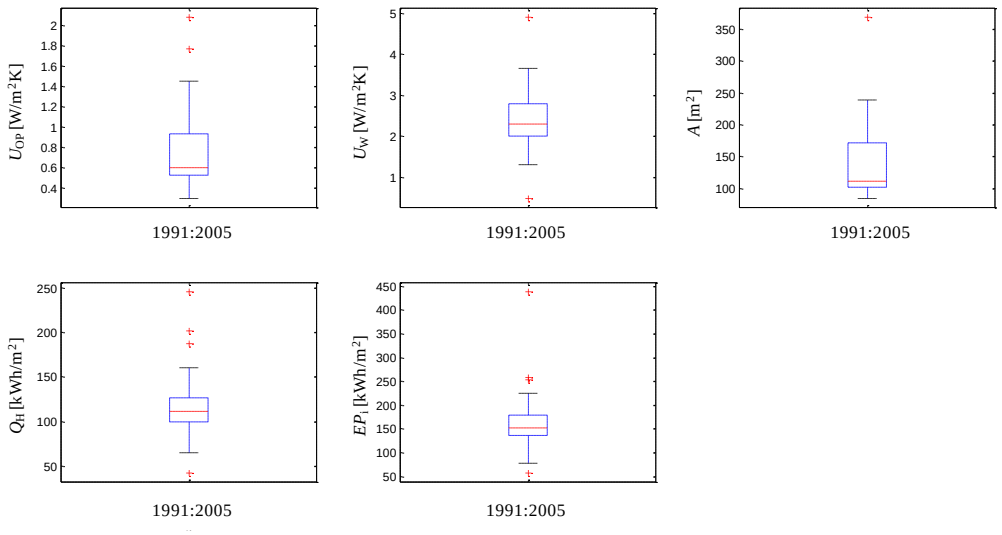

Figure 6: Boxplot of the variables considered (U_{op}, U_w, A, Q_H ,EP_i) using the data set of buildings built from 1990 to 2005.

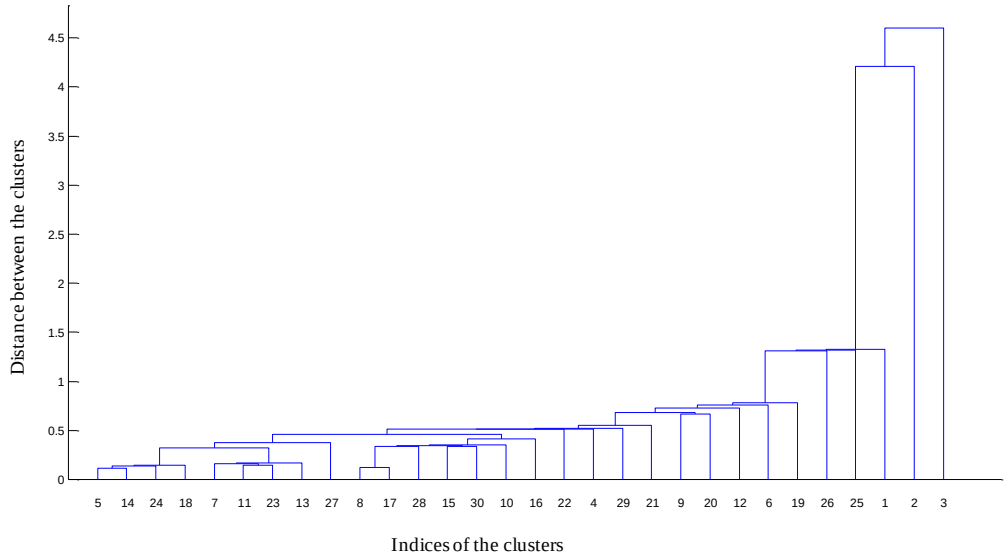

Figure 7: Terraced houses dataset of the building built from 1990 to 2005 organized in 30 clusters.

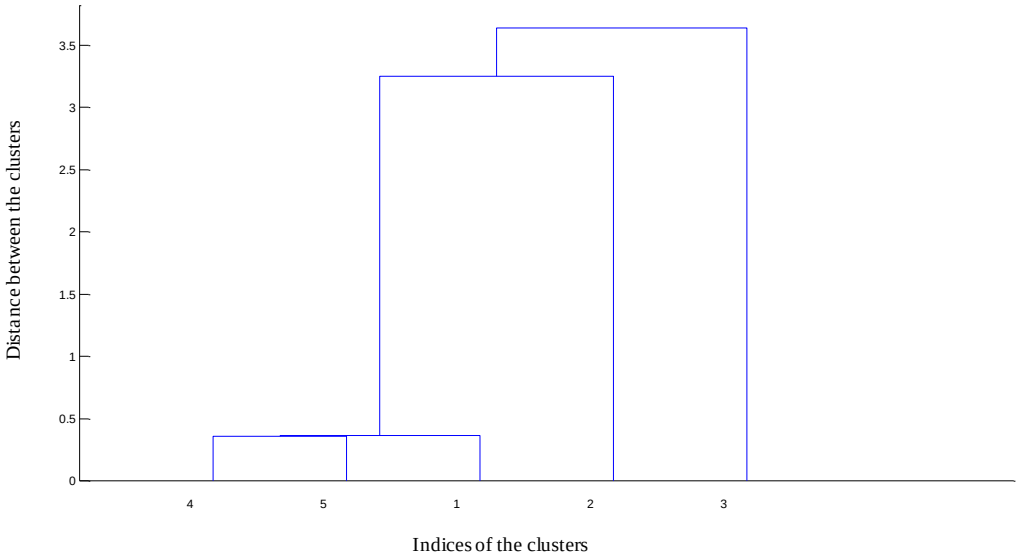

Figure 8: Terraced houses dataset of the building built from 1990 to 2005 organized in 5 clusters.