Clustering as a simplification tool for the decision-making process on building stock renovation

Mathieu Rivallain², Sergei Agapoff², Pierre Boisson², Aurélie Foucquier¹, Yunseok Lee²
¹Univ. Grenoble Alpes, CEA, LITEN, INES, Le Bourget-du-Lac, France
²CSTB, 84 avenue Jean Jaurès – Champs sur Marne, Marne la Vallée Cedex 2, France

Abstract
Representing one of the largest energy consumer, the building sector is an important issue in terms of energy consumption and climate change mitigation. In this sense, many different actions of building energy savings have been started. Among them, retrofitting measures have been taken for different building scales from the single housing units up to the district or building stock. Finding retrofitting solutions for a single house is relatively easy, whereas the building stock dimension remains a complex issue. To facilitate the work of building stock owners and local public authorities, the developed method allows identifying groups of buildings and also at the scale of the building stock to be used as input for the retrofitting actions optimization. The proposed approach consists in clustering the buildings of a given stock according to their characteristics before refurbishment - intrinsic variables as the geometry or the thermal properties – and the resulting variables as the total or heating energy consumption. Consequently, a methodology based on a clustering analysis has been implemented. It consists of successive clustering steps on a “decision space” containing the intrinsic building features and an “objective space” enclosing the energy performance. One of the main results is knowledge that the usual building classification based on climate, year of construction and type is not always sufficient and justified: compactness and height of should also be taken into account.

Introduction
The building field is one of the most energy-consuming in the world. Many improvements need to be done in order to reduce the energy consumption and one of the most efficient solutions is probably retrofitting buildings at a large scale such as the building stock. To determine the most suitable retrofitting actions, a solution frequently employed is the numerical simulations. However, at a large building scale, it begins quite limited by the fact that it is really too time-consuming and the amount of required building data is often too ambitious. Other solutions consisting in determining representative buildings are then used (Schäfer A., 2016), (Sokol Julia, 2017). Among them, clustering methods have already shown their efficiency in the understanding of a large-scale building stock for different applications from the energy assessment to retrofit activities. For that, each building must be described by a list of input attributes gathering all known buildings information as the geometrical aspects, the thermal characteristics, the climate, the occupation (Buttitta G., 2017), (Zakoveronyi A., 2017), the energy consumption (Zhao F., 2016), CO₂ emission, cost considerations or the relative change of energy consumption, CO₂ emissions and cost considerations due to retrofit measures. Obviously, all these parameters are rarely available all together but one main interest of using a clustering approach compared with numerical simulation is to be able to reduce largely the amount of input data. Moreover, considering the nature of all those attributes, one can observe that different levels of information are derived. Indeed, it is customary to have on the one hand an “objective space” containing, for instance, relative or absolute energy consumption or CO₂ emission aspects and on the other hand, a “decision space” enclosing more the geometrical, thermal or climate characteristics. The efficiency of retrofitting could be judged only considering at least one parameter of both spaces. Indeed, the retrofit measures will affect directly some parameters of the decisions space and their actions will affect consequently one or few parameters of the objectives space. This remark will take a real importance for the understanding of the buildings clusters. Actually, as we will see in the following, the ability to explain the clusters in both spaces can present some difficulties. Indeed, the main issue will be to be able to show the almost unicity of the clusters considering both the objectives and the decisions spaces in order to link parameters in one space to others in the second space.

Some studies considering the use of clustering methods in building applications have been recently published. Among them, Tardioli et al. (Tardioli G., 2018) proposed a study with the goal to assess the energy performance of a large number of buildings in Switzerland thanks to clustering methods. For that, they had a large database of more than 10 000 real mixed-use buildings containing energy performance, CO₂ emission and geometrical aspects that they finally reduced to 67 representative buildings. The clustering considered all the parameters of the database without any distinction between objective and decision spaces. Therefore, the authors showed a clear explanation of the clusters considering each decision parameters. However, regarding the objective space containing especially the energy consumption density, it appeared really less specific to each clusters leading to a difficulty to link energy performance and
geometrical aspects. Previously, Gaitani et al. (Gaitani N., 2010) presented a study consisting in the implementation of an energy classification applied to schools in Greece. They employed a building database deduced from an energy audit of 1100 schools mainly focused on heating use and considered an objective space (heating consumption) and a decision space (geometry, occupation, usage of the building, heated surface, age of the building, insulation presence, number of classrooms, number of students, schools' operating hours and the age of the heating system). A clustering method was applied to the objective space which led them to define 5 energy classes. In order to reduce the dimensionality of each cluster in the decision space, they used a PCA (principle components analysis) method. This mathematical technique consists in transforming the original parameters into a given number of new uncorrelated variables (i.e. principle components). By considering the median of each group, the authors deduced some representative buildings for each energy class. The main issue of this methodology resides in the fact that the authors did not focus on the distribution of each parameter of the decision space. Given that, it seems difficult to guarantee the unicity of each cluster considering both objective and decision spaces. In the same idea, Arambula-Lara et al. (Arambula-Lara R., 2015) proposed a study based on 60 schools energy audits in Italy to define representative buildings for retrofit measures at a large building scale. They operated the clustering on a decision space enclosing some geometrical and thermal envelope parameters. The objective space defined as the heating degree hour is not used directly in the clustering but at an early stage for the selection of variables and after the clustering as the output of a simplified model of each cluster deduced from a linear multivariable regression. This linear regression will allow in their following studies to assess the most cost-optimal retrofit measures. Representative schools are deduced from the centroid of the clusters. More recently, Deb et al. (Deb C., 2018) focused justly on this space distinction for building retrofitting applications. Their study was based on 56 energy audits in Singapore. They had both pre- and post-retrofit energy consumption and some pre-retrofit HVAC systems characteristics for each building. The final objective of the study was to determine the best set of HVAC parameters that led to the highest change in energy consumption between pre- and post-retrofit measures. Thus, the decision space enclosed the HVAC variables and the objective space was defined as the energy change between pre- and post-retrofit actions. A first step showed a large difficulty to interpret the clusters in the decision space following a clustering in the objective space. That is why, after an optimal variable selection on the decision space with a linear multivariable regression, they applied the clustering on those decision variables but observed the results in the objective space without dealing with the interpretation in the decision space. In the same manner, Geyer et al. (Geyer P., 2017) used clustering to assess in each cluster the sensitivity to different retrofit measures. The clustering was done on the objective space (post-retrofit energy consumption and CO₂ emission) calculated thanks to a factor translating the application of retrofit actions to the pre-retrofit data. Authors used only those objective variables to interpret the clusters preventing them from highlighting the link between objective and decision spaces.

The reviewed literature shows that several recent studies implemented clustering techniques to assess retrofit measures on buildings. Most of them distinguished the data between objective and decision variables but used only one of the two spaces to explain the results. Thus, the objective of this study is to understand the obtained clusters in both decision and objective spaces. To do that, we propose a successive clustering methodology embedding a degree of scientific expertise to facilitate the interpretation of clusters.

To test our methodology, three virtual building stocks generated by building energy simulation are provided. Each of them gathers a sample of 10 000 residential buildings with equal representations of individual houses, small multi-family, high-rise large multi-family and low-rise large multi-family buildings. The three building stocks are differentiated by the variable parameters considered as inputs of the building simulation.

**General methodology**

The objective of our study is to partition a large building space and to demonstrate the unicity of those partitions considering both decision and objective spaces. The methodology used in this study is based on two successive clustering steps, inside which has been added some notions of building expertise to facilitate the cluster understanding. To do that, our approach consists in using constraints at different stages of the methodology. Clustering under constraints is well known and frequently employed in the mathematical field but used very little in the building area (Brea, 2013). A first method is to declare which pairs of database’s elements should be clustered together or separately. It is usually combined with an improvement of the clustering metrics to compensate the restriction of the constraints (Basu S., 2004). A second method resides in defining constraints directly on clusters such as their number of elements, their maximal diameter or the distance between them (Brea, 2013). Considering our application, this second approach seems to be more suitable.

In the following, we will describe as generally as possible the process of the implemented methodology. The different decision and objective spaces will be defined further in the paper. In the same way, the specific choices concerning the clustering algorithm or the nature of constraints will be presented afterwards. Thus, the methodology is described in Figure 1. First, a phase of feature selection on the decision space in order to reduce the dimension of the database has been implemented. Then, after a normalization step, a clustering is applied to the reduced decision space. The number of clusters is deduced from the well-known elbow method. Thus, a
First partitioning is proposed at this stage of the methodology. Then, each cluster is cleaned by removing all the outliers concerning one specific parameter of the decision space. After that, a new normalization and clustering phase on the objective space for each cluster operates. The number of clusters is calculated from a constraint on one parameter of the objective space concerning the exceedance of its maximal standard deviation. In addition, if this constraint is not satisfied, a new clustering with only two clusters is again applied to the objective space. As a consequence, the number of clusters is not fixed.

- 1st level (I): only geometrical parameters vary;
- 2nd level (II): both geometrical and thermal envelope parameters vary;
- 3rd level (III): geometrical, thermal envelope parameters and climate vary.

Among the various outputs of the building simulation, we only considered the total energy consumption and the heating consumption. Thus, for the database I, the variability observed in the outputs will be explained only by the variation of the geometry aspect. Database II adds the variability of the thermal envelope characteristics mainly through the building year. It should be noted that the building year is directly linked with thermal coefficients of floor, ceiling, wall or window and with solar factor of the windows. Likewise, database III has the same basis as database II and adds the climate consideration. Table 1 shows the considered objective and decision spaces for databases I, II and III. Four building types are considered: individual houses, small multi-family, low-rise large multi-family, high-rise large multi-family. Climate is represented with 8 climate zones based on the weather files of the French regulation (CSTB, 2011).

The methodology has been tested on the three databases to check its robustness of the methodology.

**Table 1: Objective and Decision spaces for the three Virtual Building Stocks**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unit</th>
<th>VBS I</th>
<th>VBS II</th>
<th>VBS III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heating energy consumption</td>
<td>kWh/m²/y</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Total energy consumption</td>
<td>kWh/m²/y</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Objective space</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Floor area</td>
<td>m²</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Windows area</td>
<td>m²</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Walls area</td>
<td>m²</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Windows-wall ratio</td>
<td></td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Windows-floor ratio (WFR)</td>
<td></td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Compactness</td>
<td></td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Height</td>
<td>m</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>South windows area</td>
<td>m²</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>South WFR</td>
<td></td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Decision space</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building type (expressed as numeric field):</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1: small multi-family building</td>
<td></td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2: low-rise large multi-family building</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3: high-rise large multi-family building</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4: individual house</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Floor thermal coefficient</td>
<td>W/m²/K</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Roof thermal coefficient</td>
<td>W/m²/K</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Exterior walls thermal coefficient</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Windows thermal coefficient</td>
<td>W/m²/K</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Global heating transfer coefficient (HTC)</td>
<td>W/m²/K</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Solar factor of windows</td>
<td></td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

**Application**

**Virtual Building Stocks and feature spaces**

The study has been led on three virtual databases composed of 10,000 numerically simulated buildings. Each of them has been generated from the building energy simulation software COMETH developed by CSTB. Those three virtual building stocks (VBS) have been defined with three different levels of variable input parameters:
3621
Proceedings of the 16th IBPSA Conference
Rome, Italy, Sept. 2-4, 2019

<table>
<thead>
<tr>
<th>Year of construction</th>
<th>-</th>
<th>X</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate (expressed as numeric field)</td>
<td>-</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>1: Agen</td>
<td>-</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2: Carpentras</td>
<td>-</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3: La Rochelle</td>
<td>-</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4: Macon</td>
<td>-</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>5: Nancy</td>
<td>-</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>6: Nice</td>
<td>-</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>7: Rennes</td>
<td>-</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8: Trappes</td>
<td>-</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Feature selection

Some parameters of the decision space are highly correlated. To overcome this issue, a feature selection on the decision space takes place. The employed method is highly influenced by Deb et al (Deb C., 2018). It consists in testing several linear multivariable regressions for all possible combinations of variables. The best combination is defined as the combination with a minimal number of variables resulting in a high regression coefficient $R^2$. Thus, the number of possible combinations increases with the number of variables. Thus, for a number of variables $n_{var}$ in the combination, the following formula calculates the number of combinations:

\[
N_{comp}(n_{var}) = \frac{n!}{n_{var}!(n-n_{var})!}
\]  

Where $n$ is the total number of variables.

Constraints employed in our application

The methodology shows two different sets of constraints with a first one just after the $1^{st}$ clustering step consisting in removing the outliers considering one parameter of the decision space. In our application, we have chosen to select the building type as the decision parameter to clean. The second constraint is imposed just after the $2^{nd}$ clustering to decide whether a $3^{rd}$ clustering is needed or not. Indeed, a constraint on the maximal standard deviation of the total energy consumption is used to justify the resort to the $3^{rd}$ clustering. This maximal standard deviation has been fixed to 2.5\% of the maximal value of the total energy consumption found in the considered sub-cluster.

Data normalization

Before the clustering phases, a normalization step is needed. Thus, both the decision space for the $1^{st}$ step of clustering and the objective space for the $2^{nd}$ step of clustering are normalized considering the following formula:

\[
X_{new} = \frac{X_{old} - \min(X_{old})}{\max(X_{old}) - \min(X_{old})}
\]

Where $X_{new}$ is the normalized variable and $X_{old}$ the initial variable.

Clustering algorithm

Among many available clustering methods, we have chosen the k-means algorithm given its extreme simplicity of understanding and implementation (Xu D., 2015).

The application of the elbow method (Deb C., 2018) for the $1^{st}$ clustering step guided us to set the number of clusters to $5$.

The number of clusters for the $2^{nd}$ clustering step is deduced from the difference between maximal and minimal total energy consumption under its standard deviation found in the considered cluster (extracted from the $1^{st}$ clustering step). The objective is to minimize as far as possible the diameter of the sub-cluster in term of total energy consumption. Thus, the formula of the number of clusters $k_{2,n}$ for the $2^{nd}$ clustering step for the cluster $n$ is described as:

\[
k_{2,n} = \frac{\max(E_{tot.cluster_n}) - \min(E_{tot.cluster_n})}{\sigma_{E_{tot.cluster_n}}}
\]

Where $E_{tot.cluster_n}$ corresponds to the total energy consumption in the cluster $n$ and $\sigma_{E_{tot.cluster_n}}$ is the standard deviation of the total energy consumption in the cluster $n$.

The clustering validation is done considering the Jaccard coefficient and the RI coefficient (Xu D., 2015). Those coefficients check the data partition similarities between two clustering results on the same data. Thus, at each clustering step, ten k-means algorithms are applied and the most representative partition is stored.

Results and discussion

1$^{st}$ clustering in the decisions space

As we mentioned above, two stages can be studied for each level of database. We will first introduce results from the first clustering step. This step considers only the variables from the decision space.

For clarity we will name the clusters based on the level of database (I, II, III), their position in the $1^{st}$ step (1..5) and their position in the $2^{nd}$ step (1..m). Cluster II.3.8 is the sub-cluster 8 of cluster 3 of the database level II.

Results are presented with a boxplot visualisation generated thanks to the MATLAB software. On each box, the red central line indicates the median, and the bottom and top edges of the box indicate the 25$^{th}$ (called $q_{25}$) and 75$^{th}$ (called $q_{75}$) percentiles, respectively. The whiskers are represented in the bottom by the value $q_{25} - (q_{25} - q_{75})/2$ and in the top by the value $q_{75} - (q_{25} - q_{75})/2$. The data respectively over and under each value are indicated by red crosses.

Figure 2 shows the repartition of clusters in objectives space, whereas Figure 3 presents the results of the clustering in the reduced decision space (that is decision...
space deduced from the variables features step) with boxplot visualisation.

**Figure 2:** Boxplot (in the objective space) of the 1st clustering step applied to the reduced decision space for database I

First, we can see that the initial decision space has been reduced to 4 variables: the window-floor ratio (WFR), the compactness, the height and the building type.

Second, considering the distribution according to the type variable, we can observe in the Figure 3 that all the individual houses (corresponding the numeric field “4” on the up-right graph coordinate) are gathered in the cluster 4. Thus, all of them are split from the multi-family buildings and are contained in one specific cluster. Moreover, one can remark that by crossing the information from the type and the WFR features, it seems enough to explain the clusters of this 1st step. Thus, the following five clusters can describe this database:

1. Large low-rise and small multi-family buildings with high WFR
2. Large low-rise and small multi-family buildings with low WFR
3. Large high-rise multi-family buildings with high WFR
4. Mainly individual houses
5. Large high-rise multi-family buildings with high WFR

**Figure 3:** Boxplot (in the reduced decision space) of the 1st clustering step applied to the reduced decision space of database I

In the same way, results obtained with the 1st clustering step on database II are presented in Figure 4 and Figure 5. Among the seventeen initial features, only eight have been selected: the windows area, the compactness, the height, the south windows area, the building type, the thermal coefficient of the floor $U_{floor}$, the thermal coefficient of the windows $U_{window}$ and the heat transfer coefficient HTC. Almost the same selection has been done for the geometrical features (except the WFR that has been replaced by the windows and south windows area). All thermal envelope characteristics are directly linked with the year of construction and hence have an impact on the energy performance: older the year of construction, higher the thermal coefficients and worst the energy performance. Thereby, even the year of construction does not appear as a selected variable, it is indirectly contained in $U_{floor}$, $U_{window}$, and HTC.
Like database I, the 1st clustering step on database II separates individual houses from the multi-family buildings. However, the other predominant parameter is HTC that characterizes the energy performance of the envelope. Thus, the following description can explain each cluster:

II.1 Mainly small and low-rise multi-family buildings with a low energy performance
II.2 Large high-rise multi-family buildings
II.3 Mainly small and low-rise multi-family buildings with a high energy performance
II.4 Individual houses with a low energy performance
II.5 Individual houses with a high energy performance

Finally, the results obtained on database III are presented in Figure 6 and Figure 7. Nine variables have been selected. Among them, the same eight variables selected for database II have been retained except for $U_{floor}$.
replaced by $U_{\text{wall}}$. To those parameters is only added the climate.

On Figure 7, we can see again that the individual houses are split from the multi-family and that the HTC appears as a main parameter. Moreover, despite the boxplot visualization does not allow revealing this aspect, a third predominant feature takes place through the climate. Indeed, the following description of the clusters are registered:

III.1 Mainly small and low-rise multi-family buildings with a high energy performance for the climate of Nancy, Nice, Rennes and Trappes

III.2 Mainly small and low-rise multi-family buildings with a high energy performance for the climate of Agen, Carpentras, La Rochelle and Macon

III.3 Mainly individual houses

III.4 Mainly small and low-rise multi-family buildings with a low energy performance

III.5 Large high-rise multi-family buildings

Nevertheless, for all databases, results always seems to be less obvious in the objectives space notably due to the overlapping between clusters and a large variability of their values (cf Figure 2, Figure 4 and Figure 6). However, in an objective of energy retrofitting, a well understanding of the consumption is absolutely needed. For that, each cluster must contain a short variability of the energy consumption and must be as unique as possible. Consequently, this first step of clustering is not sufficient for our application.

2nd clustering step in the objective space: study for the database III

This 2nd clustering step is needed in order to be able to split the database considering the objective space in a more detailed manner. Let us remember that this 2nd step
can be joined with a 3rd sub-clustering step in 2 clusters if the condition on the maximal standard deviation of the total energy consumption is not satisfied. Consequently, the number of sub-clusters is varying. Moreover, a cleaning step on the building type precedes this 2nd clustering.

To lighten the results, we present a focus on database III and more precisely on clusters III.1 and III.3. Thus, Figure 8 shows the clusters distribution in the objective space for the 2nd clustering step on cluster III.1. This cluster III.1 gathers mainly small and low-rise multi-family buildings with a high-energy performance for the climate of Nancy, Nice, Rennes and Trappes. A first interest resides in the resulting unicity and the short variability of each sub-cluster considering the objective space. From this point of view, one of the main clustering goals is then reached. Let see now the understanding of each sub-cluster in the reduced decision space (cf. Figure 9). As it was mentioned before, the boxplot visualization is not suitable for representing the results according to climate feature. For readability reasons, we have chosen to summarize the results without adding another visualization plot.

A low-energy consumption (clusters III.1.3 and III.1.9) is correlated with the hot climate of Nice, a small HTC and a small compactness. Conversely, a high-energy consumption (clusters III.1.1 and III.1.7) is linked with the colder climate of Nancy, Rennes or Trappes, a large HTC, a high compactness and a small windows area.

![Figure 8: Boxplot (in the objective space) of the 2nd clustering step applied to the objective space of III.1](image)

On one hand, one can observe some sub-clusters that seems unique as the clusters III.1.7 or III.1.9. On the other hand, other clusters (like III.1.10 or III.1.12) seem similar whereas they present a real separation considering the objective space. Given the fact that it deals with a simulated database, this result can be quite surprising. Nevertheless, it can be explained by the potential interactions between the different parameters used in the clustering process. A study by implementation of non-linear surrogate models of each sub-cluster could bring some answers.

![Figure 9: Boxplot (in the reduced decision space) of the 2nd clustering step applied to the objective space of III.1](image)
Figure 10 presents the result of this 2nd clustering in the objective space of the cluster III.3 dealing with only individual houses in all climates. A large disparity is observed through the heating and the total energy consumption with amplitudes from 100 to 800 kWh/m²/year. Again, the graph shows that the methodology is well-suited to obtain final narrow distributed unique sub-clusters according to the objective space. As required by the constraint, the boxplot size decreases with the absolute value of the parameter indicating that the maximal standard deviation decreases proportionally to the maximal value of each sub-cluster. These results can be generalised to all clusters and even more to all databases.

Figure 11: Boxplot (in the reduced decision space) of the 2nd clustering step applied to the objective space of III.3

Discussion and conclusions

First, we can notice that this clustering methodology presents a strong robustness regarding the different information contained in each database.

Second, we observed a coherence in the clustering results according to the level of information contained in the different databases. Among these results, it appears that the usual building classification according to the climate, the year of construction and the type is not always sufficient and even more justified. Indeed, we saw that a first level classification can be done according to the admitted building type for all databases but most of the time, another attribute is absolutely needed as the WFR for the database I (that does not describe the energy performance variability) or the energy performance for the database II and III (more than the year of construction by the fact that it does not necessarily reflect the energy performance in case of previous retrofitting actions). As
expected, the database III, where the climate is added as input data, presents the climate as an important feature. But this first order classification is not enough to explain a building stock. Indeed, most of the time, the compactness and the information of single or multi-storey buildings (through the height) appear also as predominant attributes. More generally, there is always a combination of the whole decision and objective features allowing to interpret the sub-clusters. Consequently, a unique partition considering those two spaces is obtained leading to the unicity of each sub-cluster. However, their detailed understanding considering only the decision space can present some difficulties. Actually, the unicity of the sub-clusters in this alone decision space cannot be always guaranteed. It could be due to the fact that some parameters interactions are not considered as decision features. To overcome this issue, a better knowledge of each sub-cluster could be obtained by the implementation of surrogate models. In further researches, those surrogate models could be used as efficient and suitable tool for the determination of optimal retrofitting solutions.

Acknowledgement
The authors would like to thank the French National Research Agency (ANR) for funding this work through the REHA-PARCS research (ANR-15-CE22-0011).

References