

Exploratory Analysis of Building Stock: A Case Study for the City of Esch-sur-Alzette (Luxembourg)

Antonino Marvuglia[1](✉) and Mohamed Laib[2]

[1] Environmental Research and Innovation Department (ERIN), Luxembourg Institute of Science and Technology (LIST), L-4362 Esch-sur-Alzette, Luxembourg
antonino.marvuglia@list.lu

[2] IT for Innovative Services Department (ITIS), Luxembourg Institute of Science and Technology (LIST), L-4362 Esch-sur-Alzette, Luxembourg
mohamed.laib@list.lu

Abstract. One of the main steps in developing urban building energy models (UBEM) is the classification of the building stock according to building archetypes. Different approaches have been proposed to accomplish this task, some based on the application of clustering techniques, or a combination of expert knowledge, deterministic classification, and data driven approaches. This paper proposes the utilization of a hybrid approach where exploratory data analysis is combined with feature extraction and feature selection to support clustering. The proposed methodology was applied to the building stock of the city of Esch-sur-Alzette (Grand Duchy of Luxembourg). The used data set includes buildings' geometrical and physical characteristics, preassigned occupancy estimates, and final energy use simulated with a quasi-steady-state model. According to the variables' combination and deterministic building stock fragmentation schemes used, the number of archetypes identified varied between 12 and 89. The paper shows the potential of clustering techniques for the development of archetypes, even though this must be combined with other (deterministic) fragmentation methods because clustering alone does not allow for the differentiation of building use typologies and construction periods, both of which must be considered to characterize buildings properly.

Keywords: Building archetypes · Building energy models · Clustering

1 Introduction and State of the Art

According to the International Energy Agency (IEA), the building sector has an enormous efficiency potential that is still far from being fully leveraged [1]. In order to help policy makers and energy planners to set up priority actions for renovation at city level, comprehensive building energy models at city scale are and will be increasingly needed [2]. The concept of Urban Building Energy Models (UBEM) has now become a standard way to look at the estimation of citywide energy demand expanding from the building

© The Author(s) 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14104, pp. 374–391, 2023.
https://doi.org/10.1007/978-3-031-37105-9_25

level [3]. A comprehensive description of the approaches used in large-scale building energy models is given in the reference [3], where the four main steps of a UBEM are identified as: 1. 3D City model, 2. Archetype development, 3. Urban climate data, and 4. UBEM simulation engine. From the reference [3] it emerges that one of the key determinants of building energy models is the building stock aggregation step, whereby one constructs a model of the building stock that reflects as best as possible its characteristics. Due to space constraints, this is the step we will mostly focus on in this paper. As it is well described in [4], building stock aggregation and characterization models can be broadly divided into three categories: top-down models, statistical bottom-up models, and engineering-based or physics-based bottom-up models [5, 6]. Top-down models are macro-scale, and they do not look at individual end-uses. They treat the built environment as a single energy user and utilize historical aggregated data for their estimations. Cities are analyzed from the perspective of techno-socioeconomic drivers (e.g., by econometric equations) [7]. Bottom-up approaches consider urban attributes at the micro-scale, studying individual (or sets of) buildings. The estimation of individual end-uses is then extrapolated to a larger scale (city/regional/national). This approach relies on the availability of extensive data to gather information on uses and impacts [8, 9]. The first step to describe the building stock is the identification of the geometrical properties of buildings (geometry, shape, and geospatial positions) using 3D city models. Subsequently, non-geometrical properties of buildings, like material, system, and occupancy, are normally defined by *building archetypes*. The definition of archetypes is a bottom-up engineering modelling procedure used to classify sets of buildings according to some common characteristics so that the detailed data and model results of the building that are identified as representative of each archetype can be extrapolated to the rest of the buildings belonging to the same group [7]. The last step of UBEM is the thermal model itself using a simulation engine. Archetypes' identification is thus a major step in UBEM. However, there is still no standard method for the definition of buildings' representative archetypes [10], and archetypes development remains still one of the biggest challenges in UBEM [3]. As was clearly stated in [2], despite providing a useful initial rough classification of the building stock, the simplistic classification by building use typologies necessitates a complementary fragmentation to identify variations related to equipment and system technical specifications as well as occupant behavior.

Several approaches have been applied to identify buildings' archetypes. Most of them use statistical techniques [11], while some apply a data-driven methodology [2, 12]. In Sokol et al. [13], a Bayesian method is used to factor in occupants-related characteristics in the definition of the archetypes, using probability distributions to represent uncertain parameters for which reliable data are rarely available. Numerous studies apply cluster analysis to the building stock to identify representative building classes and improve the accuracy of energy use prediction models [10, 14]. In Tardioli et al. [15], a clustering methodology in six steps for building classification is proposed to identify representative buildings and groups of buildings characterized by similar features. This approach has the advantage of not assigning a certain weight to specific features (e.g., the energy index or total energy consumption) but instead balancing the importance of all the building characteristics (i.e., geometry, energy, and occupancy). In Borges et al. [2], building archetypes are identified by combining deterministic buildings classification (based on

characteristics like the use of typology and construction period) and clustering carried out using the R package NbClust [16] in various orders. The authors conclude that this approach allows obtaining archetypes of a higher granularity than when applying deterministic and cluster methodologies separately. In Nägeli et al. [17], a synthetic approach is used to generate realistic building stock data. In Costanzo et al. [18], instead of operating through direct archetypes identification, an approach based on using different layers of information is used, with the aim of avoiding oversimplifications. Afterward, the energy use prediction model was realized using a simulation-based approach in EnergyPlus.

The present paper proposes a methodology to achieve the buildings archetypes fragmentation that best represents buildings in terms of their expected operation energy use. The used methodology combines a deterministic method (based on the subdivision of the buildings according to their construction period and a pre-defined list of building typologies) with unsupervised clustering. The methodology is applied to a case study dealing with the city of Esch-sur-Alzette, in the South of Luxembourg.

2 Methodology

2.1 Data Description and Preparation

The proposed methodology in this paper is illustrated in Fig. 1. Each step is illustrated in the following.

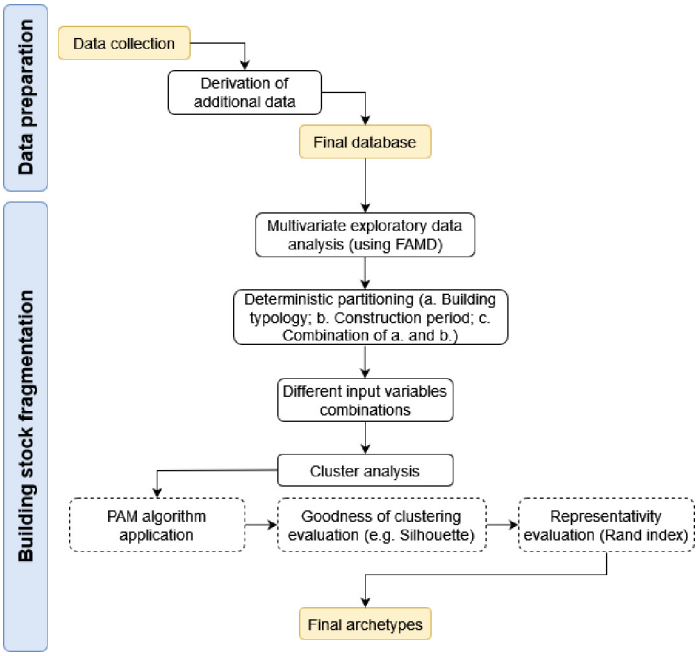


Fig. 1. Proposed hybrid methodology for archetypes' determination.

The data collection step is the same as described in [19] and [20], as here the same building stock data (geometrical characteristics, type of heating, U-values) is used. As described in [19], buildings elements and components were selected and classified according to previous studies [21, 22] and relevant standards [23]. The geospatial dataset consists of georeferenced building footprints (a georeferenced polygon for each building) and related attributes on building characteristics (year of construction, building function, and typology). The derivation of additional data consists in the calculation of geometrical characteristics like average building height (H_{avg}), building gross volume (V_{gross}), useful floor area (A_{useful}), and the area of walls delimiting the building envelope, that were obtained as described in [19], where the procedure to assign materials to each building component in each building by using the respective building type and period of construction information (and making resort to stochastic allocation in case of unknown information, such as the state of renovation) is also detailed.

As detailed in [5] and [20], the final energy use intensity (i.e., the energy used per m^2 of heated floor area) in each building was calculated using a quasi-steady-state energy demand simulation model for which the set of variables listed in Table 1 was available and which was applied to a data set containing 5400 buildings and 6594 cadastral units (see Table 2). Data variables listed in Table 1 represent our final database.

Table 1. List of variables known for each record (building) of the dataset.

Variable #	Variable name	Description
1	building_id	Building ID
2	QUARTIER	District
3	Typology	Building type
4	year_const	Construction year
5	ETPLEINS	Floors over the ground level
6	NBRHABITAN	Number of occupants
7	height_ave	Average building height (m)
8	length_w_out	Perimeter of outer walls (m)
9	surf_w_out	Surface of outer walls (m^2)
10	foot_area	Gross footprint area (m^2)
11	A_n	Heated floor area (m^2)
12	U_w	U-value of walls (W/m^2K)
13	U_r	U-value of roof (W/m^2K)
14	U_f	U-value of the ground floor (W/m^2K)
15	heating_sys	Heating system type
16	window_id	Window typology
17	$q_{E,V}$	Final energy use intensity (kWh/m^2a)

Heating system type (heating_sys) can take four values: 1. Conventional boiler of single-family-house (SFH); 2. Condensation boiler of SFH; 3. Conventional boiler of multi-family-house (MFH); 4. Condensation boiler of MFH. Window typology (window_id) can take seven values as described in Table 3, taken from [20].

Table 2. Number of buildings in the database per each building typology.

Building typology	n. of buildings	n. of cadastral units
1. Detached house (DH)	954	1159
2. Row-house (RH)	675	943
3. Mixed usage building (MX)	509	825
4. Multi-family-house (MFH)	3262	3667

Table 3. Window types.

Window type	Glazing type	Frame type	Construction period	U-value (W/m^2K)	g-value (-)
1	Single	Wood	Until 1968	5.00	0.87
2	Double	Aluminium (no thermal break)	1969–1983	4.30	0.75
3	Double	Aluminium (with thermal break)	1984–1994	3.20	0.75
4	Double	PVC	1969–1994	3.00	0.75
5	Double	Wood	Until 1994	2.70	0.75
6	Double	PVC-Aluminium	From 1995	1.90	0.60
7	Double	Wood	From 1995	1.60	0.60

2.2 Multivariate Exploratory Data Analysis

To visually explore the characteristics of the building data set, multivariate data analysis techniques have been applied in this paper. Since the dataset contained numerical and categorical features, Factor Analysis of Mixed Data (FAMD) was applied [24, 25]. The algorithm is a compromise between Principal Component Analysis (PCA) [26] and Multiple Correspondence Analysis (MCA) [27] and is known to handle well numerical and categorical features at the same time. In FAMD, each continuous variable is standardized (i.e., centred and then divided by its standard deviation), and each categorical variable is transformed into a dummy variable and divided by the square root of the proportion of objects taking the associated category [28]. Then, a PCA is applied to the resulting features (standardized for the continuous and transformed for the categorical) [29].

In this paper, FAMD is used to reduce the dimension of data to easily visualise it and gain better insight into the data structure. Moreover, FAMD's new dimensions are also used for clustering and compared to other variables' combinations. Each nominal variable has J_k levels, and the sum of all the J_k equals to J. Each nominal variable is coded using four indicator variables. For example, the four levels of the variable "heating_sys" are coded as 1000; 0100; 0010 and 0001. There are $I = 6594$ observations. We denote $\mathbf{X}$ the I $\times$ J indicator matrix (i.e., a matrix whose entries are 0 or 1). The J $\times$ J table obtained as $\mathbf{B} = \mathrm{X}^{\mathrm{T}}\mathrm{X}$ is called the *Burt matrix* associated to $\mathbf{X}$.

The proportion of variances explained by the new dimensions is displayed in the scree plot shown in Fig. 2.

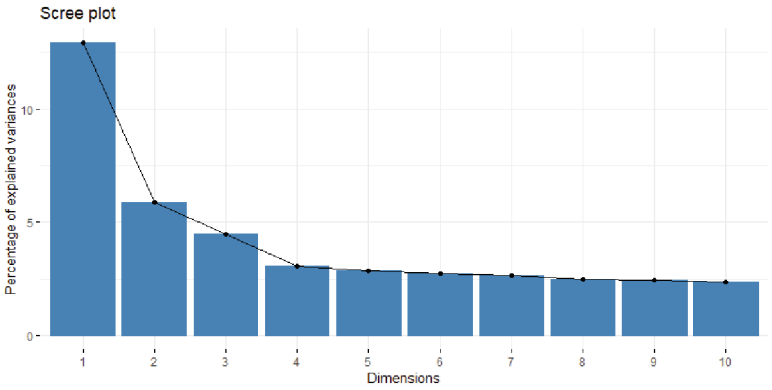


Fig. 2. Scree plot of the eigen values of the Burt matrix.

Only the first three dimensions have around 5% or higher of the explained variance. The remaining axes have an equal contribution of 2% of the variance. In this study, FAMD is applied to get quick insight and visualization of the dataset. However, in our case, the explained variance is too low (e.g., the first 10 dimensions plotted in the figure represent only 41% of the variance), and therefore more than just 10 dimensions were needed to perform the clustering described in the following sections. For the sake of readability, the data were projected into a 2-dimensional space formed by the Burt matrix's first two eigenvectors. Figure 3 shows the representation of the continuous variables on the *circle of correlations* [25], projected on the two first dimensions of FAMD.

The association between each variable is depicted on the graph. Variables that have a positive correlation are grouped. Variables with negative correlation are placed on each side of the origin (opposed quadrants). *Cos2* is a measure of how well the variables are represented on the factor map (square cosine, squared coordinates). A high *cos2* value (around 1) suggests that the variable on the principal component is well represented. In this case, the variable is situated relatively close to the correlation circle's edge. A low *cos2* value (around 0) implies that the principal components do not sufficiently describe the variable. In this case, the variable is very near the circle's centre. If a variable can be accurately described by just two principal components (Dim1 and Dim2), then the sum of the *cos2* on these two principal components is equal to one. The variables will

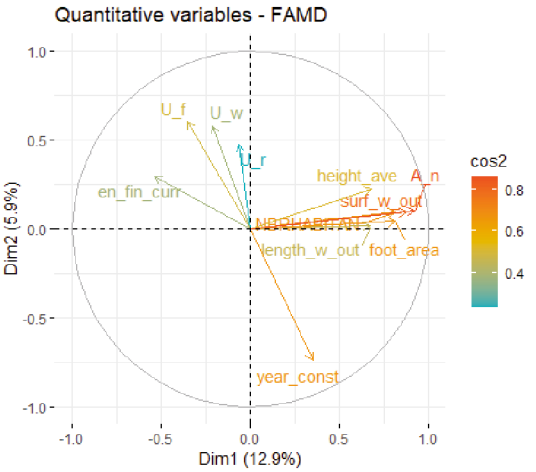


Fig. 3. Representation of the continuous features on the space spanned by the two first principal components.

be placed on the circle of correlations in this case. More than two components may be needed for some of the variables to capture the data completely. The variables in this instance are situated inside the correlation circle. In Fig. 3 the variables are coloured according to their *cos2* variables. As one can see from Fig. 3, some features are highly correlated, such as foot_area, length_w_out, and A_n.

Figure 4 shows the factor map representation of the data on the two first components.

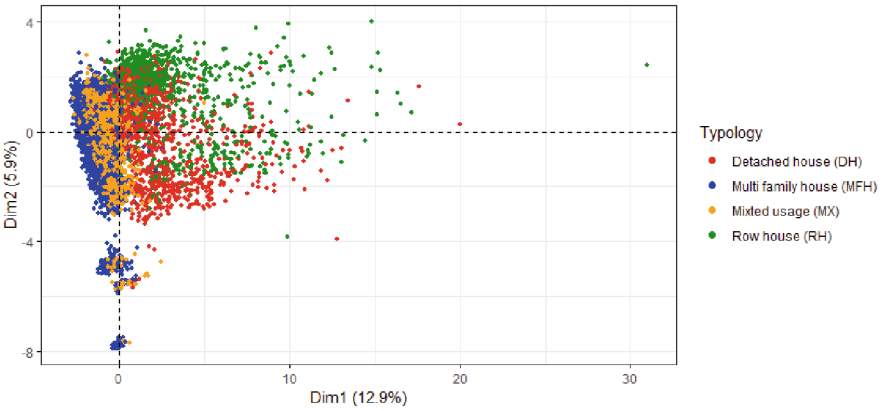


Fig. 4. Factor map representation of individuals on the two first components (coloured by building typology).

In Fig. 4 one can see that in these first two dimensions, the MFH class partially overlaps the MX class, and the DH class partially overlaps the RH class. This result is normal because MFH buildings are usually multi stores buildings, where offices or services can

be easily located, and RH ones have several similar characteristics (geometry, shape, materials, etc.) to DH.

2.3 Building Stock Fragmentation

In the building stock fragmentation step, the buildings were divided into smaller subsets using deterministic and data-driven methodologies. The aim is to obtain subsets for which the similarity is also respected in terms of final energy use intensity (energy used per m^2 of the net surface). In other words, each of the obtained buildings' subsets should contain buildings that are not only similar with respect to their physical and functional characteristics (i.e., the features used to perform the fragmentation) but for which also a similar energy use intensity can be expected. In this way, the cluster label attributed to each building can help understand the best scheme to achieve building classifications based on buildings' characteristics that best respond to their actual energy use. This last point will be better explained in Sect. 3.

The building stock fragmentation stage is divided into three steps: (1) Variables' combinations choice; (2) buildings' deterministic classification (based on building typology, on construction period, and the combination of both); (3) PAM-based clustering. Details of each of these procedures and their implementation in the case study are presented in the remainder of the paper.

We decided to use four construction periods: Period 1: Before 1900; Period 2: Between 1900 and 1950; Period 3: Between 1951 and 2000; Period 4: After 2000. These periods reflect three main "construction waves". The first one, at the beginning of the 20^{th} century, was linked to the exploitation of the iron mines and the flourishing of the steel industry, which attracted numerous workers. The second is linked with the reconstruction after World War II. A third wave took place at the end of the 20^{th} century due to the booming of the finance and consulting sector in Luxembourg (and to some extent also the European institutions) that attracted and is still attracting a considerable number of workers.

2.4 Variables' Combinations

The variables' combinations (VC) on which the following steps (deterministic classification and PAM-based clustering) were performed is based on a) expert judgment (schemes from VC1 to VC4 below), b) unsupervised features selection (scheme VC5), and c) features extraction supported by FAMD (VC6).

More specifically, we applied the following variables' combinations schemes:

VC1. All the variables (from #2 to #17) shown in Table 1;
VC2. All the variables from #2 to #16 (the final energy use intensity, $q_{E,V}$, was excluded);
VC3. All the variables, except the U-values (#12, #13 and #14) and the number of occupants (#6);
VC4. All the variables, except the U-values (#12, #13 and #14), the number of occupants (#6) and the final energy use intensity (#17);
VC5. Unsupervised feature selection on the full data set (i.e., including all the variables from #2 to #17 in Table 1) based on the space-filling concept introduced in [30];

VC6. FAMD dimensions that represent 75% of the variance in each fragmentation.

Note that the number of selected variables in VC5, as well as the number of necessary dimensions in VC6, vary for each deterministic partitioning.

The selection of a number of variables lower than the initial dimension of our data manifold is based on the rationale that, when using the entire set of variables, some partitionings contained too few data-points with respect to the number of features. The same problem was encountered in [2]. To curb this problem, we then applied features selection to reduce the number of features used.

The variables' combination choice was repeated for each of the building partitionings obtained with the three following deterministic clustering subdivisions: 1. Division by building typology (called *typ_sep* hereafter); 2. Division by period of construction (called *period_sep* hereafter); 3. Division by both building typology and period of construction (called *typ&period_sep* hereafter).

The applied unsupervised feature selection algorithm eliminates existing data redundancy and keeps only those features that include new information. The algorithm is coded in the R package '*SFtools*' [31].

Table 4 shows the list of selected features for each studied fragmentation using VC5. It is worth mentioning that some of the variables are selected for all fragmentation, such as NBRHABITAN, height_ave, $q_{E,V}$. As Table 4 shows, there are several redundant features, e.g., one can see that the variables length_w_out and surf_w_out are selected alternatively in each fragmentation scenario.

2.5 Cluster Analysis

K-means [32] is one of the most well-known clustering techniques. It is used in many fields and relies on the distance matrix of the data, which is usually calculated using the Euclidian distance metrics. However, since the data set used in this paper is characterized by a mix of categorical and numerical features, other dissimilarity measures, such as the Gower distance [33], have been considered. Furthermore, to minimise the influence of noise and outliers, a medoid-based method was required, namely the Partitioning Around Medoid (PAM) algorithm was applied [34]. The main difference between PAM and K-means is that the latter computes the mean value of the cluster (centroid) to use as a prototype vector to represent the cluster, while the former uses an existing vector (i.e., a data-point) as a representative object. For this reason, the PAM algorithm is less sensitive to the initial choice of medoids than the K-means algorithm. This characteristic of PAM contributes to further limiting the influence of noise and outliers. However, the PAM algorithm is more computationally expensive than K-means, as it requires the calculation of all pairwise distances between points at each iteration.

The R packages '*clust*' [35] and '*factoextra*' [36] have been used to carry out the analyses.

Following the exploratory data analysis result, the data is grouped into classes according to the building typology. Clustering quality has been assessed using the silhouette score [37]. If one takes any object i in the data set and denotes by A the cluster to which it has been assigned, when A contains other objects apart from i, the average distance $a(i)$ of i from all the objects within A can be calculated. If one now considers another cluster

Table 4. List of selected features for each studied fragmentation, which corresponds to VC5.

	DH	RH	MX	MFH	Before 1900	1900 to 1950	1950 to 2000	After 2000	DH & 1900 to 1950	DH & 1950 to 2000	DH & After 2000	RH & 1900 to 1950	RH & 1950 to 2000	MX & 1900 to 1950	MX & 1950 to 2000	MFH & 1900 to 1950	MFH & 1950 to 2000	MFH & After 2000
QUARTIER	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Typology																		
Year_const	✓	✓	✓	✓														
ETPLEINS	✓	✓					✓	✓	✓	✓	✓	✓	✓					
NBRHABITAN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
height_ave	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
length_w_out	✓	✓			✓	✓	✓				✓				✓			✓
surf_w_out			✓	✓				✓	✓			✓	✓	✓		✓	✓	
foot_area									✓						✓			✓
A_n			✓						✓				✓	✓		✓		✓
U_w			✓				✓		✓	✓			✓	✓	✓		✓	
U_r	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
U_f							✓		✓									
heating_sys	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
window_id	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
$q_{E,V}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

$C \neq A$ and computes the mean $d(i, c)$ of the distances from i to all the objects in C, the smallest $(b(i))$ of those distances can be selected. The silhouette width for data-point i in cluster A is defined by:

$$\begin{cases} s(i) = \frac{b(i)-a(i)}{max[a(i),b(i)]} & if\ |C_A| > 1 \\ s(i) = 0 & if\ |C_A| = 1 \end{cases} \tag{1}$$

where $|C_A|$ is the cardinality of A, i.e., the number of elements in A.

The silhouette S_A of cluster A is the average of the silhouette widths of all the objects in A. Given a certain partitioning that separates the data set into K clusters, the overall silhouette score of the partitioning is the mean of the silhouette of the clusters over all

K clusters:

$$S = \frac{1}{K} \sum_{A=1}^{K} S_A \tag{2}$$

Therefore, the higher the silhouette score, the better the clustering [37]. A good partitioning of the data set yields a silhouette score close to 1.

3 Results and Discussions

The number of clusters identified with the hybrid clustering (i.e., using first the deterministic clustering and then the PAM algorithm) per each VC and each fragmentation scheme varied between 12 and 89, with the lowest number (12) found when using the VC1 or the VC2 scheme coupled with the *typ_sep scheme* deterministic building stock fragmentation, and highest number (89) found when using the VC6 scheme coupled with *typ&period_sep.*

Figure 5 shows a heatmap of the silhouette scores of the different subgroups of buildings obtained by combining building typologies and construction periods (deterministic partitioning cases shown as column headers).

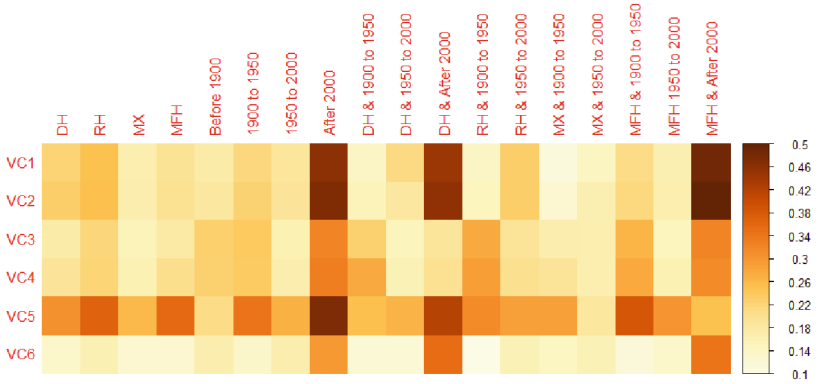


Fig. 5. Silhouette values for the partitionings obtained dividing the buildings by typology (*typ_sep*), by period of construction (*period_sep*), and by a combination of the two (*typ&period_sep*).

From the observation of Fig. 5, one can infer that the typology alone normally does not provide "optimal" clusters from the compactness and separation standpoint (measured by the Silhouette index). This is true in every case, but slightly improves when unsupervised features selection is used (VC5). In all the cases, the separation among clusters is more clear (higher values of the Silhouette index) for the buildings built after 2000. We can partially explain this by the fact that for certain old buildings (or even single dwellings) certain renovation interventions (like internal insulation or windows replacement) could have been realized without having been recorded, therefore

this information may be missing from the dataset. This is less likely for newer buildings (built after 2000).

However, as it will be shown later in Fig. 7, in our context, the first objective of clustering is using the descriptors (i.e. the variables from #2 to #16) to obtain sets of buildings that are as similar as possible in their expected energy use intensity, while separation and compactness (reflected by the Silhouette index) becomes the second objective.

Looking at the VC2 row of Fig. 5, one can observe that for this variables' combination scheme, there are three cases (buildings "After 2000", "DH & After 2000", and "MFH & After 2000") with the best partitioning (highest Silhouette values of Fig. 5), that are also reflected in terms of final energy use, i.e., when we repeat the clustering adding the variable $q_{E,V}$, (as the VC1 row of Fig. 5 exhibit nearly the same pattern in terms of Silhouette values). Among these three cases, the box plots of the final energy use of each obtained cluster showed that, in terms of clusters separation, the best case is the one of VC2 for MFH built after 2000 (Fig. 6), as the mean and median values are the best separated from one cluster to the other. Nonetheless, some overlap between the range of variation of the $q_{E,V}$ values of the different clusters is inevitable, as there will always be dwellings with different features but with the same or similar energy use intensity. As mentioned above, one reason for this is that there are buildings belonging to the same construction period but with different renovation states.

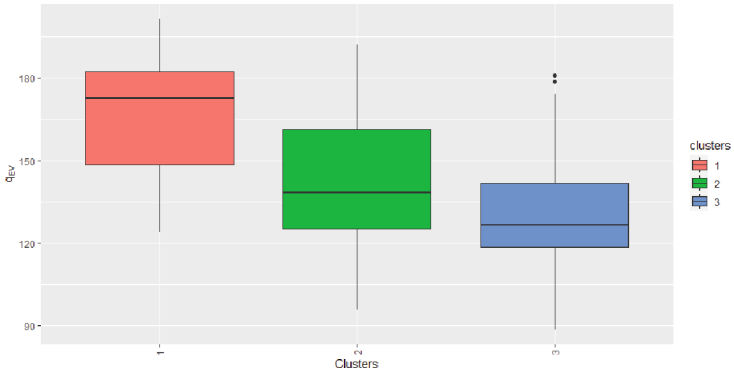


Fig. 6. Box plots of the final energy use intensity ($q_{E,V}$) of each cluster obtained with VC2 for MFH built after 2000.

Finally, the partitionings obtained using all the variables (VC1) have been compared with all the others using Rand's cluster similarity index [38]. This index takes values between 0 and 1, being 0 the scenario in which the two partitionings one is comparing have no similarities (that is, when one only consists of a single cluster and the other is composed of clusters containing single points), and 1 the case in which the partitionings are the same.

The values of the Rand similarity indices obtained are shown as numbers within each cell in Fig. 7. The figure emphasizes the similarity between VC1 and VC2, as the Rand index approaches value 1 in almost all the deterministic partitioning cases. Moreover,

the figure shows that when the U-values (variables #12, #13, and #14) and the number of occupants (variable #6) are removed from the data set (namely, the cases of VC3 and VC4) the obtained clusters are not similar (i.e., the values of the Rand index are low) to the ones obtained in the case of VC1 (i.e., when all variables are included). From this perspective, we then argue that the VC2 scheme would be the best option to obtain clusters that can reflect the expected final energy use when this latter is unknown. This conclusion is indeed not very surprising since the final energy use is calculated starting from the variables that express the physical characteristics of the buildings [5, 20].

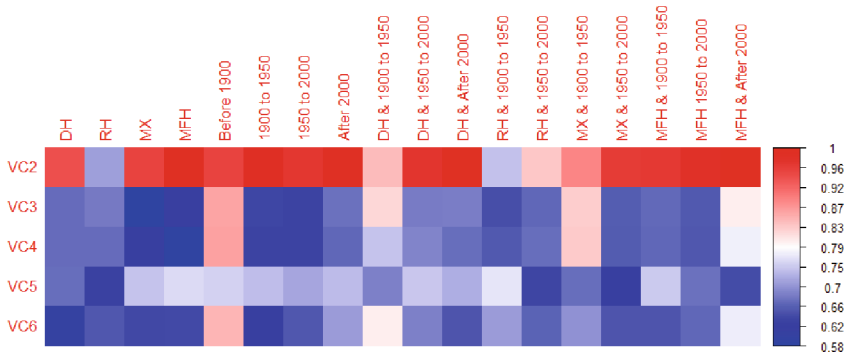


Fig. 7. Rand index values obtained by comparing the clusters resulting from VC1 (i.e., when removing the variable qE,V) with others (VC2 to VC6).

Table 5 shows the number of clusters obtained per each VC after applying the PAM algorithm only to the *typ&period_sep* scheme, which is the situation where the highest numbers of clusters are obtained (compared to *typ_sep* and *period_sep* schemes).

We can then assume that the final archetypes (last step of Fig. 1) identified by the methodology described in this paper are the 43 clusters obtained with the combination of the VC2 and *typ&period_sep* schemes (second row of Table 5).

Figure 8 shows the boxplots of q_{EV} [kwh/m^2 a] for each of the clusters identified by using the variables' combination VC2, after applying the PAM algorithm to the *typ_sep* scheme. The clusters are named using the acronym of the building typology they refer to (e.g. DH for detached houses) and the cluster number within that particular typology (e.g. DH_1 is the first of the clusters that contain detached houses).

The figure shows that the VC2 scheme, even if it does not use q_{EV} as input variable, allowed a reasonably good separation in terms of final energy use intensity (looking at the distancing among the medians and the average values of the clusters, typology by typology). The fact that a similar separation was obtained also with VC1, which, however, includes q_{EV} among the input variables, testifies that the scheme VC2 allowed to obtain a partitioning that represents building clusters (i.e. archetypes) with a reasonably good separation in terms of expected energy use intensity.

Table 5. Number of clusters per each VC scheme after applying the PAM algorithm to the *typ&period_sep* scheme.

	DH & 1900-1950	DH & 1950-2000	DH & After 2000	RH & 1900-1950	RH & 1950-2000	MX & 1900-1950	MX & 1950-2000	MFH & 1900-1950	MFH & 1950-2000	MFH & After 2000	Total n. of clusters
VC1	6	3	3	4	3	8	3	3	3	3	39
VC2	9	3	3	7	3	6	3	3	3	3	43
VC3	9	10	5	4	11	6	11	7	11	3	77
VC4	4	10	11	4	10	6	11	7	9	3	75
VC5	3	3	3	3	3	3	3	3	5	3	32
VC6	11	9	11	11	11	8	4	11	6	7	89

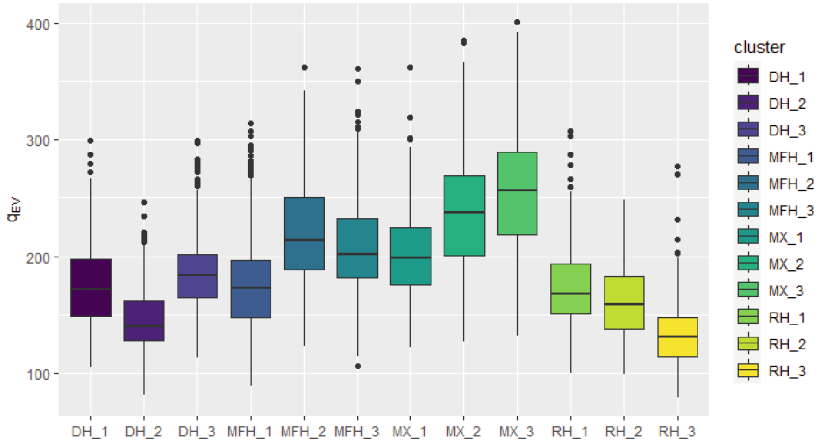


Fig. 8. Clustering applied to each deterministic split (VC2).

4 Conclusion

This research suggests a new hybrid methodology for archetype identification that combines the traditional deterministic approach with cluster analysis. The proposed approach has been applied to the building stock of the city of Esch-sur-Alzette, in the south of Luxembourg. The used building stock comprises 5400 buildings and 6594 cadastral units. The number of archetypes identified varied between 12 and 89 (according to the different schemes detailed in the paper).

The chosen archetypes are the 43 clusters obtained by applying the PAM clustering algorithm to the data set which comprises all the variables in the dataset, excluding the final energy use intensity (variables' combination scheme called VC2 in the paper), but where the buildings had been previously partitioned using the deterministic separation based on building typologies and construction period (called *typ&period_sep* in the paper).

The novelty of the proposed approach, compared to similar analyses proposed in the literature [2, 15], consists mainly in the exploration of different variables' combinations and the application of unsupervised variables selection, in addition to variables' extraction obtained with the FAMD algorithm [24].

Borges et al. [2] suggested that, when building metered energy is used as a unique variable to perform clustering, completing the cluster analysis prior to the building period fragmentation allows to better capture the patterns of energy usage and impacts the cluster analysis' outcome as little as possible. In our case, we do not use directly metered energy data, but energy data derived from a physics-based simplified model. Moreover, the clustering is performed using several variables, and the matching with the energy usage patterns is checked using the Rand index. Therefore, in our case the order of the two steps is still relevant, but to a lower extent.

We confirm the finding already highlighted by Borges et al. [2] that there is strong evidence that the use of clustering techniques has a high potential for the development of archetypes, even though this must be combined with other partitionings, because clustering alone does not allow for the differentiation of building use typologies and construction periods, both of which must be taken into account to properly characterize buildings. This fact is even more important when one considers that if the consumption patterns and heating surfaces produce similar ratios, even very dissimilar buildings in terms of design, internal gains, tenant occupation and behaviour, heating ventilation and air conditioning (HVAC) efficiency, refurbishment conditions, and energy conservation measures may have similar values of the final energy use intensity. As a result, the energy use intensity alone may be a deceptive variable for determining representative buildings for energy modelling since it is unrelated to building geometry. Clustering, on the other hand, ensures that partitions are made while considering the total range of variation in each variable. However, the use of clustering algorithms brings up other concerns, such as the sample size of the buildings in the database.

Future work will involve benefitting from the proposed hybrid approach to use buildings cluster labels (detected archetypes) as one of the input variables to inform UBEMs and validate the results of the new energy simulations using metered energy data. This, however, necessitates collecting new data.

Acknowledgments. This research was developed within the project "Continuous lifecycle assessment of buildings and districts through real-time design and operation data underpinned by semantics" (SemanticLCA), funded by the Luxembourg National Research Fund (FNR) under the Grant INTER/UKRI/19/14106247 and the UK Engineering and Physical Sciences Research Council (EPSRC) under the Grant EP/T019514/1.

References

1. IEA: Buildings - A source of enormous untapped efficiency potential. https://www.iea.org/topics/buildings. Accessed 12 May 2022
2. Borges, P., Travesset-Baro, O., Pages-Ramon, A.: Hybrid approach to representative building archetypes development for urban models – a case study in Andorra. Build. Environ. **215**, 108958 (2022). https://doi.org/10.1016/j.buildenv.2022.108958
3. Johari, F., Peronato, G., Sadeghian, P., Zhao, X., Widén, J.: Urban building energy modeling: state of the art and future prospects. Renew. Sustain. Energy Rev. **128**, 109902 (2020). https://doi.org/10.1016/j.rser.2020.109902
4. Langevin, J., et al.: Developing a common approach for classifying building stock energy models. Renew. Sustain. Energy Rev. **133**, 110276 (2020). https://doi.org/10.1016/j.rser.2020.110276
5. Mastrucci, A., Marvuglia, A., Benetto, E., Leopold, U.: A spatio-temporal life cycle assessment framework for building renovation scenarios at the urban scale. Renew. Sustain. Energy Rev. **126**, 109834 (2020). https://doi.org/10.1016/j.rser.2020.109834
6. Abbasabadi, N., Ashayeri, M.: Urban energy use modeling methods and tools: a review and an outlook. Build. Environ. **161**, 106270 (2019). https://doi.org/10.1016/j.buildenv.2019.106270
7. Swan, L.G., Ugursal, V.I.: Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. Renew. Sustain. Energy Rev. **13**, 1819–1835 (2009). https://doi.org/10.1016/j.rser.2008.09.033
8. Allegrini, J., Orehounig, K., Mavromatidis, G., Ruesch, F., Dorer, V., Evins, R.: A review of modelling approaches and tools for the simulation of district-scale energy systems. Renew. Sustain. Energy Rev. **52**, 1391–1404 (2015). https://doi.org/10.1016/j.rser.2015.07.123
9. Ferrando, M., Causone, F., Hong, T., Chen, Y.: Urban building energy modeling (UBEM) tools: a state-of-the-art review of bottom-up physics-based approaches. Sustain. Cities Soc. **62**, 102408 (2020). https://doi.org/10.1016/j.scs.2020.102408
10. Schaefer, A., Ghisi, E.: Method for obtaining reference buildings. Energy Build. **128**, 660–672 (2016). https://doi.org/10.1016/j.enbuild.2016.07.001
11. Cerezo, C., Sokol, J., AlKhaled, S., Reinhart, C., Al-Mumin, A., Hajiah, A.: Comparison of four building archetype characterization methods in urban building energy modeling (UBEM): a residential case study in Kuwait City. Energy Build. **154**, 321–334 (2017). https://doi.org/10.1016/j.enbuild.2017.08.029
12. Pasichnyi, O., Wallin, J., Kordas, O.: Data-driven building archetypes for urban building energy modelling. Energy **181**, 360–377 (2019). https://doi.org/10.1016/j.energy.2019.04.197
13. Sokol, J., Davila, C.C., Reinhart, C.F.: Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. Energy Build. **134**, 11–24 (2017). https://doi.org/10.1016/j.enbuild.2016.10.050
14. Ghiassi, N., Mahdavi, A.: Reductive bottom-up urban energy computing supported by multivariate cluster analysis. Energy Build. **144**, 372–386 (2017). https://doi.org/10.1016/j.enbuild.2017.03.004
15. Tardioli, G., Kerrigan, R., Oates, M., O'Donnell, J., Finn, D.P.: Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach. Build. Environ. **140**, 90–106 (2018). https://doi.org/10.1016/j.buildenv.2018.05.035
16. Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A.: NbClust: an R package for determining the relevant number of clusters in a data set. J. Stat. Soft. **61**, 1–36 (2014). https://doi.org/10.18637/jss.v061.i06
17. Nägeli, C., Camarasa, C., Jakob, M., Catenazzi, G., Ostermeyer, Y.: Synthetic building stocks as a way to assess the energy demand and greenhouse gas emissions of national building stocks. Energy Build. **173**, 443–460 (2018). https://doi.org/10.1016/j.enbuild.2018.05.055

18. Costanzo, V., Yao, R., Li, X., Liu, M., Li, B.: A multi-layer approach for estimating the energy use intensity on an urban scale. Cities **95**, 102467 (2019). https://doi.org/10.1016/j.cities.2019.102467
19. Mastrucci, A., Marvuglia, A., Popovici, E., Leopold, U., Benetto, E.: Geospatial characterization of building material stocks for the life cycle assessment of end-of-life scenarios at the urban scale. Resour. Conserv. Recycl. 54–66 (2017). https://doi.org/10.1016/j.resconrec.2016.07.003
20. Mastrucci, A., Pérez-López, P., Benetto, E., Leopold, U., Blanc, I.: Global sensitivity analysis as a support for the generation of simplified building stock energy models. Energy Build. **149**, 368–383 (2017). https://doi.org/10.1016/j.enbuild.2017.05.022
21. Nemry, F., et al.: Options to reduce the environmental impacts of residential buildings in the European Union—potential and costs. Energy Build. **42**, 976–984 (2010). https://doi.org/10.1016/j.enbuild.2010.01.009
22. Mastrucci, A., Popovici, E., Marvuglia, A., De Sousa, L., Benetto, E., Leopold, U.: GIS-based life cycle assessment of urban building stocks retrofitting. A bottom-up framework applied to Luxembourg. Presented at the Enviroinfo & ICT4S 2015, Copenhagen, Denmark (2015)
23. Charette, R.P., Marshall, H.E.: UNIFORMAT II elemental classification for building specifications, cost estimating, and cost analysis. NIST US Department of Commerce (1999)
24. Pagès, J.: Analyse factorielle multiple avec R. Presented at the (2013). https://doi.org/10.1051/978-2-7598-1085-7.c004
25. Kassambara, A.: Practical guide to principal component methods in R. STHDA (2017)
26. Hill, M.O., Smith, A.J.E.: Principal component analysis of taxonomic data with multi-state discrete characters. Taxon **25**, 249–255 (1976). https://doi.org/10.2307/1219449
27. Kiers, H.A.L.: Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. Psychometrika **56**, 197–212 (1991). https://doi.org/10.1007/BF02294458
28. van de Velden, M., Iodice D'Enza, A., Markos, A.: Distance-based clustering of mixed data. WIREs Comput. Stat. **11**, e1456 (2019). https://doi.org/10.1002/wics.1456
29. Vichi, M., Vicari, D., Kiers, H.A.L.: Clustering and dimension reduction for mixed variables. Behaviormetrika **46**(2), 243–269 (2019). https://doi.org/10.1007/s41237-018-0068-6
30. Laib, M., Kanevski, M.: A new algorithm for redundancy minimisation in geo-environmental data. Comput. Geosci. **133**, 104328 (2019). https://doi.org/10.1016/j.cageo.2019.104328
31. Laib, M., Kanevski, M.: SFtools: space filling based tools for data mining (2017)
32. MacQueen, J.: Some methods for classification and analysis of multivariate observations. Presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, Berkeley, Calif. (1967)
33. Gower, J.C.: A general coefficient of similarity and some of its properties. Biometrics **27**, 857–871 (1971). https://doi.org/10.2307/2528823
34. Kaufman, L., Rousseeuw, P.J.: Partitioning around medoids (Program PAM). In: Finding Groups in Data, pp. 68–125 (1990). https://doi.org/10.1002/9780470316801.ch2
35. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.: cluster: Cluster Analysis Basics and Extensions. R package version 2.1.4 (2022)
36. Kassambara, A., Mundt, F.: Factoextra: extract and visualize the results of multivariate data analyses. R Package Version 1.0.7 (2020). https://CRAN.R-project.org/package=factoextra
37. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7
38. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971). https://doi.org/10.2307/2284239

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

