Clustering Strategies for Defining Archetypes to Support Integrated Simulations of Environmental Impacts

V Gomes 1, O O C Zara 1, G M Colleto 1, M G da Silva 2

1 School of Civil Engineering, Architecture and Urbanism; University of Campinas, Brazil, vangomes@unicamp.br
2 Technology Center; Federal University of Espirito Santo, Brazil

Abstract. Life cycle assessment (LCA)’s inherent data-intensiveness hampers application to neighbourhood environmental assessments, particularly for built stock modelling. Data collection can be reduced to a manageable amount by grouping a large number of buildings into a limited set of aggregates with similar characteristics and defining exemplars (archetypes) that represent each group. LCAs would be performed for the archetypes only and their results extended to the represented buildings. As archetype definition is seldom detailed in the literature, this paper tests, and details different procedures that could enable neighbourhood LCAs. K-medoids and CLARA partition algorithms, as well as agglomerative hierarchical clustering techniques, were applied to group over 300 buildings into a limited number of clusters. A building representative of each cluster was identified to proceed to bottom-up LCA. K-medoids clustering stands out for the quality of clusters and their representatives. Restraining the maximum number of clusters to keep subsequent LCA work manageable imposes some quality loss yet allows for achieving satisfactory division results. Regardless of the clustering technique used, data was the best divided the larger the number of clusters used, for the various factors in the database depicting the studied area resulted in several possible data combinations. Although detailed representation is desirable in LCA modelling, limiting the number of variables facilitates data pre-treatment and an optimal balance should be pursued in future studies.

Keywords. neighbourhood LCA, hybrid modelling, clustering, benchmarks, data gap filling.
being assessed. Thus, finding balance between required and feasible information collection, and determining the best way to handle data quality and uncertainty must be prioritized [3].

For completeness’ sake, in process based LCA, the modelled product system must be meticulously described as a compilation of material types and respective masses flowing in and out throughout the reference service life. Hence, one can imagine that LCA application becomes progressively challenging as the aggregation level of the studied systems increase [4], for they require numerous data inputs, estimates and predictions to describe them, which are not always available or accurate enough. For LCA to become mainstream practice and offer support to quotidian decision-making, a constant quest for modelling streamlining, whilst preserving core information. Most often, this is achieved through cut-off rules that balance simplification and completeness. Though such cut-offs are often made in individual building assessments, trade-offs between the modelling level of detail and scope broadness become unavoidable for broader assessment scopes. In this sense, evolutionary approaches to improving background and foreground bottom-up archetypes can aid in balancing data collection needs for neighbourhood modelling [3], for example.

For combining simplified modelling and extended scope, the classification by archetypes is commonly used in research on new technologies [5], energy efficiency policy design [6], and urban building energy models (UBEM) [7, 8, 9, 10]. Gathering information at an urban scale to support detailed bottom-up models is a daunting task. In urban models, archetypes represent specific groups of buildings aggregated based on similar characteristics, selected according with the scope of the study at hand and with the data and computational resources available. These ‘representative buildings’ are then used to describe the overall building stock which contain them. Large building stocks have been segmented and represented by sample or archetype buildings based on similar construction year, use, type of heating system and building geometry [11] and include climate zone and main heating fuel source when validating energy efficiency measures over national stocks [12]. The K-means algorithm is often used in such cases, but presents limitations to handle qualitative, non-numerical attributes or cluster representatives expected to be actual data items, as it would be the case when screening buildings for which to conduct labour-intensive, process-based urban LCAs.

Urban LCA shares the need to simplify building stock modelling and can greatly benefit from using archetypes to streamline the life cycle inventory. Grouping several buildings into a limited set of aggregates, each of them represented by one archetype, would limit the detailed materiality survey to the archetypes only, and the corresponding inventory would be extended to the represented buildings based on their built area.

Archetype definition procedures are, however, seldom detailed in the literature [7], and information on the use of archetypes for life cycle assessment of urban areas is scarce [13]. Therefore, this paper details the machine learning-supported approach, using R coding, to determine the bare minimum number of background archetypes required for representing a selected neighbourhood’s building stock. When most of the variation in the population occurs within the clusters (i.e., a given archetype describing a relatively homogenous group of buildings) - and not between them - the expected random error is smaller. Clustering is a technique of data grouping through statistical tools and algorithms. This grouping is done automatically through unsupervised machine learning and is based solely on the information present in the variables. Over 300 buildings listed in the collected database were grouped into a limited number of clusters, making subsequent LCA manageable. Buildings with similar – even if not identical – values for the considered variables were grouped together and a building representative of each cluster was identified to proceed for further treatment for LCA purposes.

1.1 Clustering techniques
Clustering divides the entire dataset into meaningful groups (i.e., clusters) based on the patterns or groups of similar objects recognized within a data set of interest, in a way that data items similar enough are grouped together but are dissimilar across the remaining groups. Unsupervised machine learning algorithms are not guided by pre-set grouping assumptions: it only has independent variables,
i.e., no fixed target/dependent variable to predict. Gradually, the machine's algorithm learns and improves its ability to group with each run.

In a non-hierarchical cluster, the measure of distance represents the similarity. In mathematical terms, the goal is to obtain the greatest possible homogeneity within a cluster (minimum intracluster distance) and the greatest heterogeneity between clusters (maximum intercluster distance). The R language supports main non-hierarchical clustering algorithms, like K-means, K-medoids, as well as ‘Clustering LARge Applications – CLARA’, a variation of the ‘partitioning around medoids’ (PAM) algorithm tailored for large databases.

K-means partitions a set of \( n \) data points into \( k \) clusters with the lowest sum of the distances of the points to their respective centroid (intracluster squares - wcss), by minimizing the Sum of Squared Error (SSE) function [7]. An initial number of clusters is arbitrated; centroids are assumed for each cluster; the distance of each observation to the centroids is calculated; and objects that are less than a pre-established Euclidean distance from the centroid are clustered. The algorithm repeats this process until convergence, when the clusters become stable, and their composition no longer changes.

The K-medoids algorithm takes a similar line but is more robust to noises and outliers than K-means [7]: instead of representing a group by its numerical means, it uses the PAM algorithm to pick a medoid that shows most similarity to other data items in the group. As similarity is given by the distance between points, the medoid is a point \( m \) such that the sum of squared distances between points in the cluster and \( m \) is minimized, where \( m \) is an actual data point within the dataset. This calculation procedure increases the maximum or average run time (computational cost) relatively to K-means clustering [14]. That must be kept in mind, as the computation cost also increases with the number of clusters, so the right number of clusters is defined in practice by balancing quality metrics that describe the average dissimilarity between each object and the respective medoid, like the Dunn index and inertia, and the computational resources available.

To decrease the run time of using the PAM algorithm for numerous data points, the CLARA algorithm takes a small, fixed size data sample and applies the PAM algorithm to generate an optimal medoids set for the sample, at the minimum computational cost. Sampling bias is minimized by repeating the sampling and clustering procedure for a pre-set number of times.

Finally, in ‘Hierarchical Cluster Analysis’ (HCA) clustering is produced either by successively merging individual observations into most similar clusters until only one large cluster is formed (agglomeration); or by the other way around (division) [14].

K-means is the most used clustering procedure, and ubiquitous in building archetypes development and urban building energy modelling [10], for which it is considered to produce the best clustering formation [7]. However, K-means has major limitations when clusters are of different sizes, densities and when the data contains outliers [7]. Moreover, such clustering technique works fine for data items with numerical attributes, for the cluster centroid is synthetically represented by the mean of the data points, and not by one of them. Hence, centroids are not interpretable data points. This particularity becomes an issue in cases where data items have qualitative attributes or when calculating an average, theoretical data point of the set as its representative makes little practical sense. In such cases - like in ours, each cluster representative is expected to be an actual data item, i.e.: we want to find building representatives of certain groups of existing buildings, for which to conduct process-based, bottom-up LCA. K-medoids partitioning then offers a better fit. Hierarchical, Density-based and K-Medoids clustering algorithms have been also tested for UBEM applications [7].

### 1.2 Clustering quality metrics

Clustering requires that computational methods identify distance or (dis)similarity between each pair of observations and produce a ‘dissimilarity’ or ‘distance’ matrix. Most methods within R coding – as well as in other languages - only calculate distance for data with a single variable type, typically numerical. The ‘daisy()’ function (cluster pack) uses the Gower metric to calculate the distance matrix () of non-numerical columns or mixed data. The Gower index combines different types of variables
and processes them according to their own mathematical type. It first calculates the partial similarity of each descriptor between the objects of interest, and then normalizes the data so that there is no difference across the different magnitudes. The algorithm is described in [15], but it basically has the following logic: to compare 2 people, A and B, in an X1 variable, one should first check whether a comparison is possible - that is, whether there are X1 measures for both person A and person B. If such comparison is possible, a score is given to how similar these two people are. For the variable “Nationality”, for instance, one can assign the value 1 if both persons are Brazilian. The general similarity score between person A and person B is the sum of all similarities of individual variables divided by the total of possible comparisons, i.e., the number of variables for which data are available for both individuals.

When using the PAM algorithm, the silhouette method determines best the relevant number of clusters [16]. The silhouette value compares how similar an object is to its own cluster relative to other clusters, with a high value indicating that the object is well combined within its cluster but poorly combined with neighbouring ones.

2. Methods
Our method comprised three main steps:

- Visual/image analysis of the built stock for data collection, using Google Inc. applications;
- Data pre-treatment, i.e., cleaning and normalizing the data table describing architectural and constructive features before running clustering routines; and
- Identification of possible patterns and rules of association for archetype definition, by cluster analysis by similarity

These steps’ output would indicate a representative building for each cluster, for which bottom up LCAs would be subsequently performed, by construction element. Cut-off rules can further limit data requirements and improve modelling but detailing the bottom up LCA is off scope in this paper, which focuses on archetype definition and the clustering procedures.

The archetype definition began with preliminary characterization of the built stock. Located in the municipality of Campinas, the study area comprises 601,012 m² of gross floor area (GFA) for mixed uses, such as: administration, education (classrooms), research (laboratories, workshops), health facilities (hospitals, clinics), libraries, restaurants, cultural facilities, sport facilities, general services, day care centre, school, bank agencies, squares, public spaces and more.

For simplification’s sake, facilities smaller than the university standard building (i.e., < 500 m² GFA) were cut-off. The remaining 321 buildings were characterized by overall architectural features (GFA, number of floors, plan shape) and specific visual/image analysis of the most relevant elements for embodied impact assessment, i.e., structural frame and envelope (roof and façades/windows). Relevant information for future operational energy assessment were also registered, like the construction completion year, shading solutions and windows-to-wall ratio.

Clustering can be performed in several ways, using statistical packages or coding in e.g., Matlab, Visual Basic, R or Python languages. This paper uses the RStudio statistical software and specific functions within the ‘cluster’ and ‘Rtsne’ packages. Unsupervised learning algorithms - K-medoids and CLARA - and hierarchical clustering were considered for selecting the most suitable technique and number of clusters for our study. Results differ little between agglomerative and divisive hierarchical clustering, so we opted for the former method, for being the most often practiced. As HCA requires no predefined number of clusters, any method can be used to evaluate the relevant number of clusters. We then consistently used the silhouette method throughout our investigation, for it is the dominant approach for this purpose within clustering studies.

Working with the least number of clusters – hence, of representative buildings – minimizes the LCA modelling intensity. Thus, a 10-cluster range (~3% of the over 300-building database at hand) was arbitrarily defined as manageable for the subsequent LCA task. Clustering quality metrics comprised
distance (inertia) and the silhouette value, and the Gower coefficient was used to calculate the distance between the mixed values (factors and numbers).

3. Results and discussion

For K-medoids, the largest the number of clusters the best is the division (Figure 1). For the pre-set 10-cluster range, the best clustering option using this method is to partition into nine clusters, as it has the largest silhouette width within the desired 10-cluster range. Figure 2 shows the resulting K-medoids partition into 9 clusters.

![Figure 1. Silhouette analysis for defining the number of clusters by K-medoids partitioning (up to 30 clusters): within the 10-cluster range, a nine-cluster partitioning had the largest silhouette width.](image1)

![Figure 2. K-medoids partitioning into nine clusters.](image2)

To reduce the number of dimensions (variables) under analysis, we applied the Principal Component Analysis (PCA), a dimensional reduction algorithm that uses linear combination of all variables to generate two representative ones (X and Y axes in Figure 2). Data dispersion in Figure 2 demonstrates the difficulty of generating a small number of high-quality clusters from a multiple-factor database. Also, approximation of data points in different clusters reflects (some degree of) similar characteristics in data points clustered differently.
CLARA is a variant of the PAM algorithm, so – as for PAM - the silhouette method is used to determine the best number of clusters. As CLARA uses several different samples to calculate the medoids rather than the whole database, the literature recommends using the average silhouette width.

Figure 3 shows that the data is the best divided the largest the number of clusters used (in this case, 38).

Within the established 10-cluster range, a four-cluster partitioning had the largest silhouette width. The resulting CLARA partition (Figure 4) shows four more well-separated clusters than in the K-medoids grouping, but with less dense data points concentration within clusters, confirming good dissimilarity among clusters, which however encompassed somewhat dissimilar data.

![Figure 3. Silhouette analysis for defining the number of clusters by CLARA partitioning (up to 50 clusters): within the 10-cluster range, a four-cluster partitioning had the largest silhouette width.](image)

PAM algorithm also identified representative items within the hierarchically clustered buildings. R language provides eight agglomerative hierarchical clustering methods. The closest the correlation between the cophenetic distances calculated by each of those methods and the Gower index distance is to 1, the most accurate is the distance representation (i.e., heights) within the dendrogram offered by the hierarchical clustering method. The ‘average’ agglomeration method shows the highest distance correlation (0.92) and was selected to generate the dendrogram. The intense branching in Figure 6 confirms that the data is best divided into a large number of clusters (68, as shown in Figure 5). Within the established 10-cluster range, again a four-cluster partitioning had the largest silhouette width. One of the clusters contains a single observation, which can be common when clustering dendrograms, but forms more scattered clusters (Figure 7).
Figure 4. CLARA partitioning into four clusters.

Figure 5. Silhouette analysis for defining the number of clusters by agglomerative HCA (up to 70 clusters): within the 10-cluster range, a four-cluster partitioning has the largest silhouette width.

All clustering techniques tested led to partitioning into numerous clusters. When observing the practicality limitation imposed (number of clusters below 10), both CLARA and the hierarchical clustering pointed towards four clusters. This number was however considered too low by the authors, who preferred to use the K-medoids clustering, which offered high clusters homogeneity and quality of cluster representatives, within the pre-set archetype range.
4. **Conclusions and outlook**

Inherent data-intensiveness hampers LCA application to neighbourhood environmental assessments. But grouping a large number of buildings of the built stock into a limited set of aggregates with similar characteristics, each of them represented by an archetype, would significantly reduce data collection and modelling effort.

The K-means is the most used clustering algorithm, including for building stock aggregation. However, when screening building candidates to conduct labour-intensive process-based LCAs, we were interested in finding cluster representatives that were one of the data items (buildings) within the cluster - so that we could use actual bill of quantities for the inventory - instead of synthetic averaged information. Hence the K-medoids algorithm fits best this purpose and indeed produced homogeneous clusters and good cluster representatives in our study.

K-medoids method is computationally more expensive, and the computation cost also increases with the number of clusters, so the right number of clusters is ultimately defined by balancing quality metrics and the computational resources available or - as also in our case - by practical reasons. Restraining the maximum number of clusters (up to 10, in this case) to keep subsequent LCA work manageable imposes
some quality loss yet allowing satisfactory division results and smoothly running the clustering algorithms.

Observing the pre-set practical limit of 10 clusters, the silhouette analyses for CLARA and agglomerative hierarchical clustering indicated classification into four clusters, but not as compact as expected. K-medoids clustering into nine groups improved cluster formation, and was the approach ultimately chosen for this clustering problem.

Regardless of the clustering technique used, data was consistently best divided into a considerable number of clusters, for the various features used to depict the existing building stock in the data table resulted in too many possible data combinations. The K-medoids calculations ran well for our 300+ building sample, after considerable effort was put on data pre-treatment, to clean noisy data and reduce the number of dimensions. Still, while essential for LCA purposes – strategies to further reduce the variables/building features to consider should be explored in future studies.

Acknowledgments
The authors thank the National Council for Scientific and Technological Development (CNPq) [GMC, PhD process # 140412/2018-2; VGS, productivity grant 313409/2021-8] and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) [OOCZ, PhD scholarship, Finance Code 001] for their generous funding; and Antonio Felipe for the coding and analytic support.

References
[1] Beloin-Saint-Pierre D et al. 2017 A review of urban metabolism studies to identify key methodological choices for future harmonization and implementation J. Clean. Prod. 163 223-40
[2] Loiseau E et al. 2012 Environmental assessment of a territory: An overview of existing tools and methods J. Environ. Manage 112: 213–225
[3] Zara OOC et al. 2020 Balancing data requirement and modelling quality in neighbourhood life cycle assessments IOP Conf. Ser.: Earth Environ. Sci. 588 042030 DOI 10.1088/1755-1315/588/4/042030
[4] Gomes V and Pulgrossi LM 2020 When part is too little: cutoff rules’ influence on LCA application to whole-building studies 11th Windsor Conference: Resilient comfort in a heating world Proc. (Cumberland Lodge: London Metropolitan University & Heriot Watt University) ed. S Roaf, F Nicol and W Finlayson
[5] Dall’o’ G, Galante A and Torri M 2012 A methodology for the energy performance classification of residential building stock on an urban scale Energy Build 48:211–9
[6] Ballarini I et al. 2014 Use of reference buildings to assess the energy saving potentials of the residential building stock: The experience of TABULA project Energy Policy 68:273–84.
[7] Ali U et al 2018 Comparative analysis of machine learning algorithms for building archetypes development in urban building energy modeling. Building Performance Modeling Conference and SimBuild Proc. (Chicago: ASHRAE and IBPSA-USA)
[8] De Jaeger I, Reynders G, Callebaut C and Saelens D 2020 A building clustering approach for urban energy simulations Energy Build 208 109671
[9] Famuyibo AA, Duffy A and Strachan P 2012 Developing archetypes for domestic dwellings - An Irish case study Energy Build 50:150–7
[10] Tardioli G et al. 2018 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
[11] Österbring M et al 2016 A differentiated description of building-stocks for a georeferenced urban bottom-up building-stock model Energy and Buildings 120: 78–84. http://dx.doi.org/10.1016/j.enbuild.2016.03.060
[12] Mata E, Sasic Kalagasidis A, Johnsson F 2014 Building-Stock Aggregation through Archetype Buildings: France, Germany, Spain and the UK Building and Environment 81:270-282. http://dx.doi.org/10.1016/j.buildenv.2014.06.013

[13] Lavagna M et al. 2018 Benchmarks for environmental impact of housing in Europe: Definition of archetypes and LCA of the residential building stock Build Environ 145:260–75

[14] Kaufman L, Rousseeuw PJ 2009 Finding Groups in Data: an Introduction to Cluster Analysis (Hoboken: John Wiley & Sons) vol 344.

[15] Gower J C 1971 A general coefficient of similarity and some of its properties Biometrics 27:857-871. http://dx.doi.org/10.2307/2528823

[16] Rousseeuw P J 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis Journal of computational and applied mathematics 20:53–65.