Acta Polytechnica CTU Proceedings **38**:169–175, 2022
https://doi.org/10.14311/APP.2022.38.0169

EXPLORING MACHINE LEARNING-BASED ARCHETYPES FOR URBAN LIFE CYCLE MODELING (UBIM)

Giseli Mary Colleto, Vanessa Gomes*

University of Campinas, School of Civil Engineering and Architecture and Urbanism, Rua Saturnino de Brito, n° 224, Cidade Universitária Zeferino Vaz. CEP: 13083-889 – Campinas – São Paulo, Brazil

* corresponding author: vangomes@unicamp.br

Abstract. Urban analyses demand simplifications that balance modelling level of detail and scope broadness. Thus, classification by archetypes is a promising methodological approach. Such an approach is common for energy studies but rarely applied for Life Cycle Assessment (LCA) purposes. When archetypes are used in urban LCA, they generally result from previous studies for classification and characterization according to parameters that directly affect the operational energy performance of buildings. This paper tackles two research questions: i) Is it appropriate to aggregate building stocks based on operational energy (OE) variables when life cycle impacts are investigated? ii) When integrated LCA (OE + embodied impacts) is pursued, would variables describing both interests simultaneously result in better representation than using operational energy-based clustering to predict embodied impacts and vice versa? Thus, we aim to confirm that, combining variables that govern OE and embodied impacts offers a better result than using OE to predict materials groupings, even if some adherence is lost relatively to single-objective clustering. Clustering experiments were carried out for the campus of the University of Campinas, Brazil. After unsupervised k-medoid (PAM) grouping, the data were submitted to a supervised learning (neural networks) classification method. Generated confusion matrices demonstrate how adherent the clustering is when considering one interest to predict the other in three situations. Results indicate that an operational energy-driven archetype fails to represent buildings from the embodied impacts viewpoint, and that merging operational energy and embodied impact variables would better support integrated life cycle impact predictions.

Keywords: Archetypes, building stock aggregation, clustering, LCA, life cycle impacts, urban modelling.

1. Introduction

Considering the growth projected for cities over the next decades, achieving established energy use goals become indispensable to mitigate urban environmental impacts. Estimating and understanding environmental impacts, both operational and embodied, are essential to assist in evaluating and monitoring performance and to support strategic planning and the formulation of energy and environmental policies for sustainable development.

From the environmental charges induced by the construction sector, studies are increasingly focusing on simultaneously reducing Operational Energy (OE) and embodied impacts [1, 2] as a decision-making ais. To avoid that only impacts during the operational phase of buildings are considered [3], Life Cycle Assessment (LCA) and energy simulations must be integrated [4], to consider environmental burden over a building's entire life cycle.

Life cycle assessments at individual building scale have already been extensively covered by literature [5]. However, the need to mitigate local and global impacts to adapt to the limitations of available resources [6] makes environmental assessment at the urban scale increasingly necessary for sustainable planning. Assessments and simulations of the urban tissue are even more complex as they face a series of challenges imposed by the scale, complexity and intensity of data needed to adequately describe urban layers and components [7].

However, if we consider the city as a set of individual buildings, the high level of detail in the analysis of segregated structures requires intensive use of financial, temporal, technological and human resources. Thus, to carry out these analyses, the complexity at the urban scale allied to the lack of detailed information about the built stock demands strategies to simplify the built stock modelling, to balance level of detail and scope broadness. Thus, archetype-based approach, for instance, abstracts existing built stock into a number of reference buildings. Construction archetypes are representations (real sample or theoretical average) of groups of buildings aggregated according to similar characteristics, defined according to the scope of the study at hand. For combining simplified modelling and extended scope, the classification by archetypes is commonly used in research on new technologies [8] and energy efficiency policy design [9]. So far, their use has increased yet been focused on operational energy studies at the urban scale, but building archetypes emerge as a promising methodological approach to extend assessment scope beyond energy use.

The development of archetypes for energy studies is based on data directly related to the operational energy of buildings. According to the grouping criterion, the representative buildings vary, reflecting the extrapolation of characteristics to the other buildings of the group that he represents and, consequently, the consistency of the simulations for the urban scale. Specific studies assessing the environmental impact of built stock at various scales point out LCA studies applying the archetype approach use the same variables as energy studies (use typology, construction typology, year/period of construction), either because they aim only at assessing operational impacts, or because they are derived from energy studies, or simply because of data availability [5]. However, an archetype for energy modeling may not be suitable for analyzing the flow of materials through the built stock [10], and representative buildings of energy perspective are not necessarily also representative from the materiality viewpoint. So, for integrated simulations of environmental assessments, the challenge of generating construction archetypes that integrate information on operational and embodied impacts arises.

This paper raises two research questions:

(i) Is it appropriate to aggregate building stocks based on operational energy (OE) variables when life cycle impacts are investigated?

(ii) When integrated life cycle (OE + embodied impacts) assessment is pursued, would variables describing both interests simultaneously result in better representation?

Thus, we aim to confirm that, at the representative building identification stage of the building archetype creation process, combining variables that govern OE and embodied impacts offers a better result than using OE to predict materials groupings and vice versa, even if some adherence is lost relatively to single-objective clustering.

2. Materials and methods

To achieve the above-mentioned goal, our method comprised two main steps. Firstly, based on a systematic literature review on archetype development methods for energy studies [11], the most used clustering parameters were identified according to the relevance of each study (operational energy and LCA). Then, cluster analyses to define archetypes were carried out for a selected case study, considering three different aggregation criteria: by energy-relevant variables; by materiality-relevant variables; and by combining energy and materiality variables simultaneously.

The case study comprised the campus of the State University of Campinas (UNICAMP), located in the municipality of Campinas, São Paulo, Brazil. With 3 893 958 m^2 total area and about 50 000 daily visitors before the COVID19 pandemic, the study area is comparable to a small town. The campus comprises 598 429 m^2 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 center, schools, bank agencies, squares, public spaces and more.

For simplification's sake, buildings smaller than the university standard building (i.e., $< 500\,\mathrm{m}^2$ GFA) and missing data points were cutoff, so that 226 buildings remained in the final sample. Geometric and non-geometric data for detailing the sample was surveyed from several sources: GIS, field visits, image analysis, data provided by administrative sectors, and results of previous studies conducted by the research group [12].

From the operational energy standpoint, data collection focused on the most used parameters for EO clustering found in the literature:

(i) Year of Construction;

(ii) Construction Typology (isolated, semi-detached, attached);

(iii) Use Typology (Administrative, Education, Hospital, Restaurant, Bank, Laboratory, Staff, Library, Museum, Gymnasium).

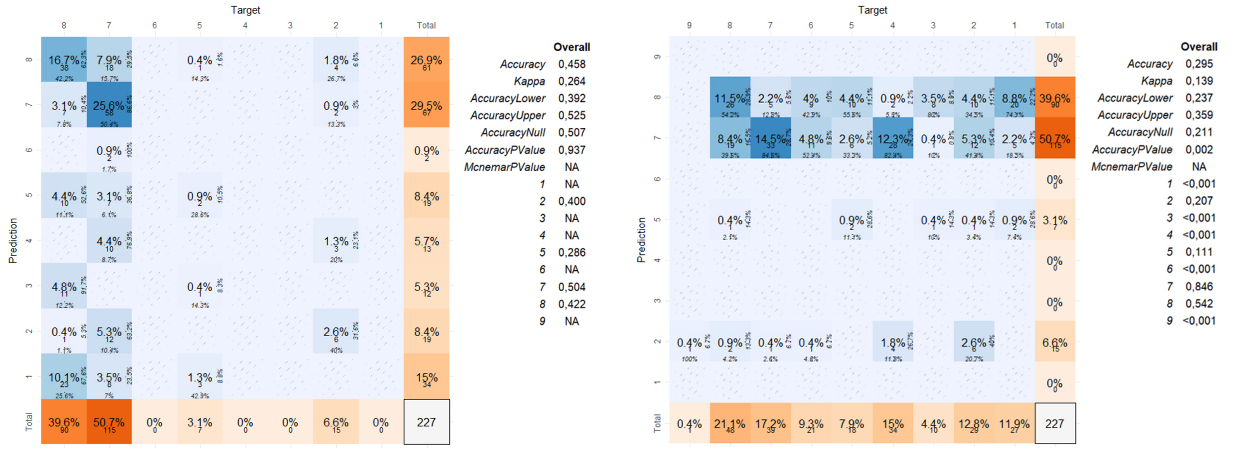
From the materiality perspective, aggregation focused on the structure (Reinforcing Concrete, Metallic, Structural concrete masonry, Pre-fabricated Concrete, Wood) and on the envelope (External Wall: Solid ceramic brick, Structural concrete masonry, Ceramic brick, Pre-fabricated Concrete, Wood; and Roofing: Metal, Concrete slab, Fiber cement, Ceramic, Polycarbonate), which govern a substantial portion of building impacts, as demonstrated by life cycle assessment studies [1]. The Total Area and Number of Floors are relevant inputs for both perspectives.

Based on a previous study on the application of different unsupervised machine learning clustering methods to the same neighborhood [7], the k-medoids / Partitioning Around Medoids (PAM) algorithm was chosen to generate a preset number of 9 clusters. In k-medoids algorithms, the groups are defined as subsets of data points that are closest to their representative elements – the medoids – which are real objects from the dataset. The identified medoids serve as the basis for developing archetypes to represent each building group. The k-medoids method is more computationally expensive (i.e. has longer processing time) than gravity center calculation, but is suitable if the groups are spherical, each medoid occupies a more central position in the group and can handle any kind of attribute (quantitative and qualitative).

Three unsupervised learning clustering scenarios were performed: operational energy-based (OE), materiality-based (LCA) and integrated clustering (OE+LCA) (Table 1). After being processed to generate clusters for each scenario and identify respective medoids, the data were submitted to neural net-

Scenario	Parameters used for clustering	Method	Cluster nr.
1	OE-related parameters: Year of Construction, Construction Typology, Use Typology, Total Area and Number of Floors	Unsupervised machine learning: K-medoids / Partitioning Around Medoids (PAM)	9
2	Materiality-related parameters: Structure, External Wall, Roof Tile, Total Area and Number of Floors		
3	Integrated clustering considering simultaneously OE parameters and materiality-related parameters: Year of Construction, Construction Typology, Use Typology, Structure, External Wall, Roof Tile, Total Area and Number of Floors		

Table 1. Description of the clustering scenarios.



(A) . Ref: OE | Pred: embodied impact. Accuracy = 45,8 %.

(B) . Ref: OE | Pred: integrated LC impacts. Accuracy = 29.5 %

Figure 1. Confusion matrices relating operational energy (OE) parameters (reference) to predicted embodied (A) and to integrated life cycle impacts (UBiM) (B) and respective model accuracies.

works, a supervised learning classification method, using the CARET package in R language [13], and the NNET method. The statistical procedures were performed using R packages “tidyverse”, for data manipulation [14]; “ggpubr”, for graph elements [15]; “cluster”, for Partitioning Around Medoids (PAM) algorithm – k-medoid [16]; “caret”, for neural networks (supervised learning) [17] and “factoextra”, for generating cluster graphs [18].

Finally, confusion matrices were generated to demonstrate how adherent the clustering is when considering one interest to predict the other in three situations:

(i) considering energy grouping to predict materiality aspects (Figure 1a);

(ii) considering energy to predict the integrated scenario (OE+LCA) (Figure 1b), and

(iii) considering materiality aspects to predict the integrated scenario (Figure 2).

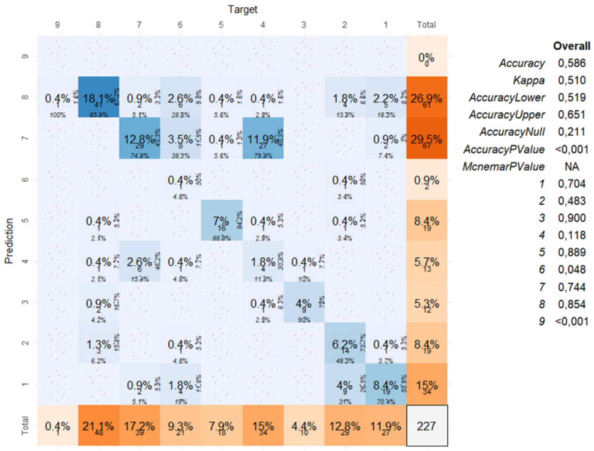


Figure 2. Confusion matrix relating embodied impact parameters (reference) to predicted integrated life cycle impacts (UBiM). Model accuracy = 58.6 %.

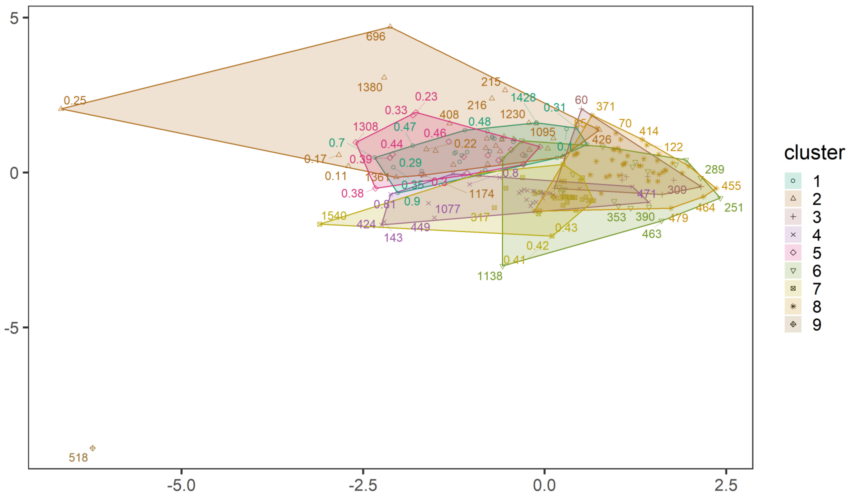


Figure 3. Clustering based on integrated life cycle (OE and embodied) impacts parameters (scenario 3). Numbers indicate the buildings IDs.

Such matrices reveal how confused the model is between the predicted classes versus the actual outcomes, and highlights instances in which one class is confused for the other, providing insightful information regarding a model's accuracy. The matrix columns are the "true" classes (reference), while the rows are "predicted" classes. The main diagonal shows the cases where the model is correct.

3. Results and Discussion

3.1. Literature review and theoretical deepening

Several efforts are being made to improve the process of creating building archetypes. De Jagger et al. [19] summarized the most common variables within 32 studies and Colleto and Silva [11] synthetized 21 publications that highlighted in the archetype development process. Table 2 gathers the main parameters from those two studies, and the relevance of each variable for energy and embodied impact assessment.

Other parameters have been identified as not directly relevant for LCA, despite their high relevance for energy:

- Building geometry:
  - ▷ Building footprint: 4 occurrences in [19], 5 in [11];
  - ▷ Compactness ratio: 1 occurrence in [19], 2 in [11];
  - ▷ Heated volume or floor area: 2 occurrences in [19] 3 in [11];
  - ▷ Total loss area: 2 occurrences in [19], 1 in [11];
  - ▷ Density of internal thermal mass: 2 occurrences in [19];
  - ▷ Ground elevation: 1 case in [11];
  - ▷ Roof type/shape: 1 case in [19], 1 in [11];
  - ▷ Aspect ratio: 1 occurrence in [11];
  - ▷ Envelope shape: 1 occurrence in [11];
  - ▷ Floor loss area / Geometry / Loss-to-floor area ratio: 1 occurrence in [19] each;
  - ▷ Position of internal thermal mass: 1 occurrence in [3];
  - ▷ Total area to net area ratio: 1 occurrence in [11];
- Building occupancy:
  - ▷ Occupancy: 3 occurrences in [19];
  - ▷ Use of ground floor: 2 occurrences in [19];
- Thermal quality:
  - ▷ U-value: 5 occurrences in [19], 3 in [11];
  - ▷ Maintenance state: 3 occurrences in [19];
  - ▷ Air tightness: 2 occurrences in [19];
  - ▷ Construction method: 1 occurrence in [19];
- Building installations:
  - ▷ Fuel typed used: 6 occurrences in [19], 3 in [11];
  - ▷ HVAC system 6 occurrences in [10], 6 [11];
  - ▷ DHW cylinder insulation thickness / Measured energy demand: 2 occurrences in [11] each; and
- Building environment:
  - ▷ Density of urban area / Exposure: 1 occurrence in [19] each.

The parameter analysis in Table 2 corroborates [10] regarding the unsuitability of an archetype for energy modeling to support materials flow analyses, by reflecting on the relationship of OE variables on materiality data needed for embodied impact assessment. The neural network classification and resulting confusion matrices enabled to quantify such inadequacy for the selected case study.

3.2. Sample clustering

The three unsupervised clustering scenarios resulted in quite distinct results. For example, in the integrated clustering scenario 3 (Figure 3), the 9 clusters (C1 – C9) were identified in spherical groups, containing a number of elements (n): C1 = 27 n, C2 = 29 n, C3 = 10 n, C4 = 34 n, C5 = 18 n, C6 = 21 n, C7 = 39 n,

	Parameters	Relevance to embodied impacts	Relevance to OE
Building geometry	Construction typology (attached, semi-detached, independent/isolated, continuous) (15) [19] (9) [11]	**Low** – Influences the building materials proportions, but its relation to material types and masses is limited • Influences the quantities of materials (e.g., attached buildings share walls)	**High** – It defines important characteristics for the building's operational energy performance due to differences in solar incidence, wind, thermal mass etc.
	Façade-area ratio (1) [19]	• Proportion of façade materials or roof materials out of the total amount of building materials and their respective impacts	**High** – Essential geometric data for operational energy performance calculations and simulations
	Roof-area ratio (1) [19]		
	Building height (3) [19] (5) [11]	**Medium** – Data support materials quantity surveying • Multiplier for material estimation in vertical elements (walls, columns etc)	
	Window area (3) [19] (2) [11]	• Supports window material and – jointly with WWR – wall area estimates	
	Window-to-wall ratio (WWR) (4) [19] (3) [11]	• Relative proportion of materials in windows and walls and their respective impacts	
	Floor area (5) [11]	• relevant for quantity surveying	
	Nr. of dwellings (buildings) / rooms (residences) (2) [11]	• multiplier for the bill of materials of a building floor; • materials estimation of internal elements	
	Number of floors / stories (5) [19] (7) [11]	**High** – Data support materials quantity surveying • Multiplier for the bill of materials for a multistorey building	
	Total area (8) [19]	• Multiplier for the total bill of materials	
	Facade materials (1) [11]	• Essential data to support the bill of quantities calculations	**High** – Non-geometric data essential for OE calculations/ simulations
Building occupancy	Use typology / Final use (14) [19] (13) [11]	**Low** – Does not standardize building materials used in construction. Can be compensated for by correction factors that represent, for example, important differences in building systems	**High** – equipment and use intensity/schedule are proxies for energy consumption patterns of different end use typologies.
Building age	Construction year/period (20) [19] (19) [11]	**Medium** – Allows retrieving the probable physical/material configuration from the implemented building regulation. In places without such regulations, it only points out building trends.	**High** – allows retrieving the probable physical/material configuration from building regulations. If absent, utility for inferring building trends for archetypes development is limited.
	Year of building's last renovation (1) [11]	**High** – Enables potential for retrofit/renovation analyses and maintenance, which directly influence a buildings' materiality.	**High** – enables retrofit potential analysis
Building environment	Climatic zone (12) [19] (3) [11]	**High** – Enables appropriate materials selection according to raw material sources/transportation, and energy matrix.	**High** – Non-geometric data essential for calculations and simulations
	Location (3) [19]		

Table 2. Analysis of the main parameters used for building archetype development. Numbers within parentheses indicate the occurrences in the reviewed literature.

C8 = 48 n and C9 = 1 n. Cluster C9 contains only one building (campus hospital), for its exceptional total built area. For each cluster, the medoid (M1 = ID1367; M2 = ID1076; M3 = ID0.12; M4 = ID301; M5 = ID0.46; M6 = ID396; M7 = ID99; M8 = ID400; M9 = ID518) is the representative building to support subsequent archetype development.

In confusion matrices (Figures 1 and 2), the columns are the "true" classes (reference), while the rows are "predicted" classes (in our case, building clusters). As the main diagonal shows the cases for which the model is correct, the ideal model shows a highly populated main diagonal, which sums up the highest sample percentage (in our case, 226 buildings). All other cells indicate prediction errors, so the "dirtiest" a confusion matrix is, the lowest its accuracy.

Despite the accuracy achieved, applying OE parameters to predict materiality grouping is the least suitable: due to the lack of data correlations, few clusters had positive results and most of them were not even statistically processed (indicated as "N/A" in Figure 1a). Contrastingly, the model's accuracy roughly doubles if life cycle aspects prediction is based on materiality (Figure 2) instead of in OE aspects (Figure 1b): 133 of the 226 sampled buildings would be correctly grouped, and clusters C8 (41/48 buildings) and C7 (29/39 buildings) were the most accurately predicted. This reinforces our assumption that, regardless of its often use, using energy-based clustering and then performing LCA for embodied impacts estimation of built stocks is not recommendable, and variables simultaneously representing both OE and materiality interests would result in better grouping than by using energy clustering for material grouping prediction and vice versa.

4. Conclusions

Balancing the most appropriate parameters for clustering vs. data collection and computational cost is a major challenge for creating representative archetypes to support built stock aggregation modelling. Despite improvement opportunities, the exploratory simulations shown support the assumption that an integrated clustering step combining operational and embodied impacts variables offers better outcomes than single-objective clustering followed by complementary simulation or LCA of representative buildings. Studies are underway to advance in the archetype development process and data simulation for predicting benchmarks for some use typologies and should be generally reproducible for varied contexts. Outcomes are expected to guide existing built stock characterization and future data collection.

Acknowledgements

The authors thank the National Council for Scientific and Technological Development (CNPq) for the PhD [GMC, process # 140412/2018-2] and scientific productivity [VGS, #306048/2018-3] grants.

References

[1] H. Birgisdottir, A. Moncaster, A. H. Wiberg, et al. IEA EBC annex 57 'evaluation of embodied energy and CO_{2eq} for building construction'. *Energy and Buildings* **154**:72–80, 2017. https://doi.org/10.1016/j.enbuild.2017.08.030.

[2] L. F. Cabeza, L. Rincón, V. Vilariño, et al. Life cycle assessment (LCA) and life cycle energy analysis (LCEA) of buildings and the building sector: A review. *Renewable and Sustainable Energy Reviews* **29**:394–416, 2014. https://doi.org/10.1016/j.rser.2013.08.037.

[3] M. Österbring, É. Mata, F. Jonsson, H. Wallbaum. A methodology for spatial modelling of energy and resource use of buildings in urbanized areas. In *SB14 Barcelona*. 2014. http://publications.lib.chalmers.se/records/fulltext/205110/local_205110.pdf.

[4] K. Allacker, V. Castellani, G. Baldinelli, et al. Energy simulation and LCA for macro-scale analysis of eco-innovations in the housing stock. *The International Journal of Life Cycle Assessment* **24**(6):989–1008, 2019. https://doi.org/10.1007/s11367-018-1548-3.

[5] A. Mastrucci, A. Marvuglia, U. Leopold, E. Benetto. Life Cycle Assessment of building stocks from urban to transnational scales: A review. *Renewable and Sustainable Energy Reviews* **74**:316–332, 2017. https://doi.org/10.1016/j.rser.2017.02.060.

[6] T. M. Baynes, T. Wiedmann. General approaches for assessing urban environmental sustainability. *Current Opinion in Environmental Sustainability* **4**(4):458–464, 2012. Human settlements and industrial systems, https://doi.org/10.1016/j.cosust.2012.09.003.

[7] V. Gomes, O. O. C. Zara, G. Colleto, et al. Archetype generation for neighbourhood lifecycle assessment. In *CB2021 Parkstad*. 2021.

[8] C. Cerezo, J. Sokol, S. AlKhaled, et al. Comparison of four building archetype characterization methods in urban building energy modeling (UBEM): A residential case study in Kuwait City. *Energy and Buildings* **154**:321–334, 2017. https://doi.org/10.1016/j.enbuild.2017.08.029.

[9] C. S. Monteiro, A. Pina, C. Cerezo, et al. The use of multi-detail building archetypes in urban energy modelling. *Energy Procedia* **111**:817–825, 2017. 8th International Conference on Sustainability in Energy and Buildings, SEB-16, 11-13 September 2016, Turin, Italy, https://doi.org/10.1016/j.egypro.2017.03.244.

[10] International Energy Agency, Energy Conservation in Buildings and Community Systems. LCA methods for buildings. Annex 31 Energy-related environmental impact of buildings, 2001. [2022-02-16], http://www.iea-ebc.org/Data/publications/EBC_Annex_31_LCA_Methods_for_Buildings.pdf.

[11] G. M. Colleto, V. G. Silva. Review of archetype development strategies for energy assessment at urban scale. In *CB2021 Parkstad*. 2021.

[12] O. Zara, G. Guimarães, M. Zibetti, et al. Balancing data requirement and modelling quality in neighbourhood life cycle assessments. *IOP Conference Series: Earth and Environmental Science* **588**(4):042030, 2020. https://doi.org/10.1088/1755-1315/588/4/042030.

[13] R core team. R: A language and environment for statistical computing, 2021. [2022-02-16], `https://www.R-project.org/`.

[14] H. Wickham, M. Averick, J. Bryan, et al. Welcome to the tidyverse. *Journal of Open Source Software* **4**(43):1686, 2019. `https://doi.org/10.21105/joss.01686`.

[15] A. Kassambara. ggpubr: 'ggplot2' based publication ready plots. R package version 0.4.0., 2020. [2022-02-16], `https://CRAN.R-project.org/package=ggpubr`.

[16] M. Maechler, P. Rousseeuw, A. Struyf, et al. Cluster analysis basics and extensions. R package version 2.1.2., 2021. [2022-02-16], `https://CRAN.R-project.org/package=cluster`.

[17] M. Kuhn. Classification and regression training. R package version 6.0-90, 2021. [2022-02-16], `https://CRAN.R-project.org/package=caret`.

[18] A. Kassambara, F. Mundt. Factoextra: Extract and visualize the results of multivariate data analyses, 2020. [2022-02-16], `https://CRAN.R-project.org/package=factoextra`.

[19] I. De Jaeger, G. Reynders, C. Callebaut, D. Saelens. A building clustering approach for urban energy simulations. *Energy and Buildings* **208**:109671, 2020. `https://doi.org/10.1016/j.enbuild.2019.109671`.