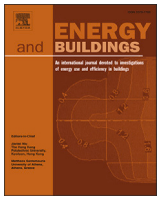
Energy & Buildings 311 (2024) 114111

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild



Building stock modelling using k-prototype: A framework for representative archetype development

Mousa Alrasheed *, Monjur Mourshed

Cardiff University, School of Engineering, The Parade, Cardiff, CF24 3AA, United Kingdom

ARTICLE INFO

Keywords:
Building stock model
Archetype
k-prototype
Representativeness
Clustering
Minimum segmentation frequency (MSF)

ABSTRACT

Building stock modelling often employs clustering techniques on the segmented stock data to identify representative archetypes, enabling cost-effective analyses while retaining the diversity and characteristics of the overall stock. However, the effectiveness of these archetypes in representing the original stock attributes remains under-explored, a factor essential for meaningful interpretations of the model outputs. This study investigated the influence of segmentation level, clustering evaluation metric and variable count on archetype representativeness by applying the *k*-prototype algorithm to the English Housing Survey data. Pre-clustering segmentation significantly influenced the outcomes, leading to the introduction of "minimum segmentation frequency" (MSF) to retain feature diversity in the segmented data. Sensitivity analysis revealed that lower MSF values improve building stock representation, while the choice of clustering evaluation metrics influences the optimal number of archetypes for a given MSF. The Davies-Bouldin index consistently identified more archetypes and achieved higher representativeness than the Calinski-Harabasz and Silhouette indices. A comprehensive archetype development framework was devised considering the influencing factors such as geographical and temporal scales, computational cost and research focus. This framework serves as a flexible guide for developing representative archetypes in future building stock modelling studies.

1. Introduction

Building stock modelling plays a vital role in the development and testing of solutions and policies for improving energy efficiency [1–3], reducing greenhouse gas emissions [4–6], adapting to climate change by reducing overheating risks [7–9], assessing the effects of building envelope modifications on indoor air quality [10] and optimising resource usage [11–13] for a resilient built environment. The modelling process can be classified into two main approaches: the "one-to-one" method, which involves modelling every building within the study area, covering a broad spectrum of its diverse geometric and construction features, and the "archetype-based" method, which focuses on modelling only a representative subset of buildings. The former method has seen increased adoption in recent years, especially for smaller geographies with fewer buildings. This is primarily due to the declining cost of computation, and the advancement and increased availability of accessible building simulation tools [14]. However, their implementation remains challenging because "one-to-one" modelling requires significant efforts in terms of human and financial resources [15]. On the other hand, in situations where many buildings need to be assessed using detailed and resource-intensive modelling approaches, developing building archetypes based on statistical analyses of a representative sample is a more feasible alternative to "one-to-one" modelling.

Each archetype embodies a range of characteristics of a particular segment of the building stock, which are often simulated to evaluate performance across a range of similar buildings while managing computational costs. Therefore, archetype-based stock modelling provides a pragmatic and time-efficient approach [16,17] while ensuring that the outcomes obtained adequately reflect the original larger set of buildings and are well-suited for their intended applications, spanning from district, and urban energy and environmental modelling to national stock modelling.

Building archetypes are primarily developed through a three-step process involving data preprocessing, segmentation and clustering. First, the building stock dataset is analysed to identify relevant features that are significant in the study context, typically using statistical methods [18]. Significant features are sometimes transformed depending on the nature of their distribution and the presence of outliers to improve clustering effectiveness [19]. Second, the selected subset is

* Corresponding author.
E-mail addresses: Al-RasheedM1@cardiff.ac.uk (M. Alrasheed), MourshedM@cardiff.ac.uk (M. Mourshed).

https://doi.org/10.1016/j.enbuild.2024.114111
Received 31 October 2023; Received in revised form 28 February 2024; Accepted 24 March 2024
Available online 29 March 2024

segmented into homogeneous groups typically based on geography and building characteristics such as age and type. Third, clustering methods are applied on each of the segmented subsets to further divide the sub-population into clusters of building archetypes with similar attributes. The level of segmentation and the selection of clustering technique depends on several factors such as the scope of the analysis, the availability and type of the variables required for modelling, and the computational complexity of the building model.

While the field of building stock modelling has seen various advancements, there remains a notable gap in the existing literature regarding a comprehensive understanding of how methodological choices affect representativeness. Representativeness can be defined as the similarity in the distribution of relevant variables between the archetypes and original building stock data, measured by comparing the total dwelling count across various variables. This relates to how well the building archetypes represent the features of the building stock, which is essential for interpreting research results effectively. Moreover, the current state-of-the-art lacks a guiding framework for developing representative building archetypes through the clustering of features relevant to specific research contexts. The absence of a systematic and adaptable framework in this regard can lead to oversimplified archetypes, potentially compromising their usability and the accuracy of simulation results. While increased archetype complexity does not guarantee improved simulation accuracy, it is important to avoid oversimplification which might overlook significant details pertinent to the building stock. An overly detailed archetype may introduce further challenges, often without a corresponding improvement in the accuracy of predictions. The key lies in finding the ideal level of detail that captures the essential characteristics needed to achieve the objectives of the study.

To overcome the existing gaps in the literature, this research explores how segmentation level, clustering evaluation metric and variable count influence the number and distribution of archetypes using the *k*-prototype clustering algorithm. Relevant variables for clustering were identified from the 2020 English Housing Survey (EHS) dataset [20] using multiple linear regression. Then, a pre-clustering partitioning strategy termed, “minimum segmentation frequency” (MSF), was introduced to retain feature diversity in the segmented data. The sensitivity of archetype representativeness to various segmentation levels, clustering evaluation metrics and variable counts was subsequently investigated. Based on the outcomes of the sensitivity analysis, a framework for representative archetype development was proposed. The framework aims to guide users in developing archetypes considering geographical and temporal scales, research focus and associated computational cost for simulation, thereby enhancing the usability and relevance of the resulting building stock model. The necessity for such a framework is underscored by the evolving nature of building stocks and increasing complexity of modelling requirements.

2. Previous works

Building archetypes have become fundamental in building stock models, serving as representative buildings that address a wide range of research objectives, from mitigating overheating risks [21,22] to reducing greenhouse gas emissions [4–6]. The characteristics of available data and specific study objectives influence the development of archetypes, emphasising the need for a systematic approach and a robust understanding of the complexities involved in their formulation.

2.1. Modelling principles

Various approaches have been adopted to develop building archetypes, and can be broadly classified into: bottom-up and top-down. Research using the bottom-up approach relies on the engineering models of the identified archetypes, the results from which are then extrapolated to the building stock using weightings. On the other hand, top-down approach often relies on statistical modelling techniques, applied on the aggregated stock data, focusing on identifying broad patterns without necessarily categorising the building stock into archetypes. Engineering or physics-based modelling also varies depending on the underlying modelling principles such as steady-state, quasi-steady-state and dynamic [23]. Dynamic energy models enable the investigation of detailed scenarios, but are often associated with high computational costs— limiting the number of archetypes to less than a hundred [24–26]. In contrast, steady-state models are less resource-intensive, and can accommodate more archetypes [27,28], potentially with increased building stock representation. However, it is essential to acknowledge that not all building stock models use dynamic models, a decision that depends on the research focus and available resources.

2.2. Segmentation and clustering

The selection of an appropriate clustering algorithm is an important consideration, guided by both the research objectives and nature of the dataset [29]. Past research on archetype development has typically employed unsupervised machine learning techniques for clustering building stock data. Two of the most popular techniques reported in the literature are *k*-means [30–36] and *k*-medoids [32,33,37], which are partitional clustering techniques that assign each instance to exactly one of *k* mutually exclusive partitions. The former method is not well-suited to concurrently handle building data comprising both numerical and categorical variables. While *k*-medoids clustering handles heterogeneous data, its computational efficiency can be affected for large datasets [38], typical of building stocks. Similarly, hierarchical clustering also has been used to develop building archetypes [30,32]. On the other hand, the *k*-prototype algorithm can simultaneously manage categorical and numerical data, and although recognised as one of the most effective methods for handling heterogeneous data [39] such as building stocks, its application for developing building archetypes remains largely unexplored.

Pre-clustering segmentation[1] or partitioning of the primary dataset has been found to capture the diversity of the building stock better than without [31], thus enhancing the representativeness of resulting archetypes. Borges et al. [34] used a deterministic method followed by *k*-means clustering to investigate the intricacies of Andorra’s building stock. Similarly, Ali et al. [31] first developed typologies through segmentation and subsequently employed *k*-means for clustering on the Irish building stock. However, the *k*-means algorithm, while effective in many cases, can fail to handle categorical variables and not account for aspects such as the total number of dwellings of each archetype. Tardioli et al. [32] explored multiple clustering algorithms on segmented subsets but did not consider *k*-prototype clustering. Furthermore, Borges et al. [34] and Tardioli et al. [32] did not partition the segmented typologies into smaller datasets, which could have potentially enhanced the stock representation achieved through clustering.

2.3. Representativeness

A key factor when determining the representation of a building stock is the number of archetypes, which typically ranges from two to several thousand in previous works. Lechtenböhmer and Schüring [40]

[1] The terms segmentation and clustering are sometimes used synonymously in the literature as they both involve grouping of cases, but differences exist between them. In the context of archetype development, segmentation is an analysis-driven process that involves grouping cases into segments based on the scope and objectives of the study. Segmentation is usually applied on the primary dataset before clustering. On the other hand, clustering is a statistical technique that uses machine learning algorithms to group cases or data points into clusters based on their similarities. One of the key differences between segmentation and clustering is that segmentation is typically driven by human knowledge and expertise, while clustering is driven by machine learning algorithms.

used only two archetypes, one 120 m^2 single/two family and one 1457 m^2 large apartment building, to represent the European Union (EU) residential building stock. The authors acknowledged significant uncertainties arising from their choice of two archetypes. Nevertheless, the research offered approximate estimations of the potential, suitability and cost associated with upgrading the EU building stock. Portella [41] developed a building stock model for France using 45 non-residential and 54 residential archetypes. The final energy demand was estimated at 435.5 TWh/year for the residential and 179.4 TWh/year for the non-residential sectors, which were 1.1-7.4% lower than the official statistics. Famuyibo et al. [18] developed 13 archetypes to represent approximately 65% of the Irish housing stock, indicating that some studies might choose fewer archetypes even if they offer limited representation of the building stock.

Research by Molina et al. [42] on the residential building stock of Chile demonstrated that a set of 496 archetypes represented the entire stock comprising 6.5 million dwellings while 90 of these archetypes represented 95% of the stock. The difference of 406 archetypes between the two thresholds indicated the presence of a large number of outlier archetypes. A χ^2 analysis in the same research revealed that the return on representativeness diminishes with increasing number of archetypes. The suitable number of archetypes was found to be dependent on the level of detail in the information sources and the desired outcomes or research questions. These findings highlight the variability in archetype selection, often influenced by different levels of segmentation, indicating the importance of methodological decisions in building stock modelling research.

In larger, national-scale investigations, a broader range of archetypes is needed to account for the diversity in building characteristics and regional disparities, as highlighted in previous studies [27,28,43]. On the other hand, studies focused on a geographically limited, district-scale scope can achieve satisfactory representation of the building stock with fewer archetypes, owing to the more uniform set of characteristics in such areas [13,30,36]. However, it's important to note that even studies of the same geographic scale may require varying range of archetypes [44,45], reflecting the diverse goals and subtleties of each research. While geographic scale often serves as a determinant for the number of archetypes, the distinct objectives of each study can further influence their selection, emphasising the complexities involved in archetype development.

The archetypes developed by Ballarini and Corrado [46] utilised averaged values of building features based on heating systems and construction typologies. This approach can be helpful in contexts with limited data but may fail to account for the variability that exists in the building stock. A more granular approach, such as clustering each typology subset, could leverage the available data more effectively than average values, leading to archetypes that reflect stock diversity more closely. Using information theory and cluster analysis, Geraldi and Ghisi's [47] advanced approach attempts to overcome such limitations by incorporating real-world parameter variability into their archetypes. Nevertheless, the approach requires extensive computational resources and depends on subjective decision factors such as spatial configurations.

3. Methodology

The proposed four-step methodology for developing building archetypes, illustrated with an example for better contextualisation, is shown in Fig. 1. The process begins with the identification of variables frequently employed in previous research. Subsequent steps involve the identification, selection, cleaning, cross-referencing and transformation of pertinent datasets. Key variables are then identified via regression analysis followed by the partitioing of the primary dataset into frequency-based subsets. A clustering algorithm is subsequently applied to each subset to generate representative archetypes. A case number is assigned to each archetype through the algorithm to determine the distribution of each archetype within the EHS. These case numbers are then used to link the archetypes to their corresponding cases in the EHS. This allows for obtaining the total dwelling count each archetype represents. The innovation of this methodology resides in the incorporation of MSF during segmentation, procedures for data transformation and variable selection, and the adoption of a suitable clustering evaluation metric and variable count— collectively contributing to an enhanced representation. The involved steps are discussed in detail in the following sub-sections.

3.1. Data preparation

Primary datasets for archetype development usually consist of both numerical and categorical variables. For instance, geometric attributes such as floor area are numerical, whereas technical features such as heating systems and fuel types fall into the categorical category. The type of variable not only affects the selection of clustering algorithm and evaluation metrics but also impacts domain-specific modelling at the end of the clustering process. Data preparation and transformation are, therefore, important steps for archetype development.

The EHS [20] was selected as the primary dataset for this study due to various considerations: (a) the comprehensiveness and reliability of the dataset, (b) its status as one of the most extensively studied building stock and (c) the opportunity it offers for a more substantial contextualisation of research findings. The EHS is a national survey of the energy efficiency and condition of housing, and people's housing circumstances in England [48]. The survey is commissioned by the Department for Levelling Up, Housing and Communities (DLUHC) and has been run since 1967. Data is collected via a household interview and a physical inspection of a sample of properties by a qualified professional. The independent categorical variables from the EHS were transformed into binary variables to satisfy the prerequisites for multiple linear regression [49,50]. Additionally, clustering outputs can be biased by skewed distributions and outliers [51], thus, scalarising data prior to clustering is essential to provide uniform weighting. Given its robust performance with various clustering methods [32], the Min-Max scalarisation was used to convert the floor area variable into a common scale ranging from 0 to 1 to improve the clustering performance.

3.2. Variable selection

Variables used in previous archetype development works are shown in Fig. 2. Dwelling type and age emerged as the most frequently used variables. Some household characteristic variables such as household size [52] and tenure [35], have seen comparatively limited utilisation. This might be attributed to the prevalent assumption of standardised occupancy profiles for dwelling archetypes, which consequently leads to excluding these variables from clustering algorithms. Variables such as ventilation systems [31] and the thickness of domestic hot water cylinders [18], are rarely included, primarily because they are absent from most datasets. The omission of ventilation systems in analyses is often due to the limited variation in building stock. For instance, the majority of the UK homes rely on natural ventilation. Additionally, modelling challenges associated with ventilation [53], may also contribute to its exclusion. An important variable implemented is energy data [34,35,52], which associates each dwelling type with its total energy consumption to establish a suitable benchmark.

A multiple linear regression model was used to examine the energy efficiency of the building stock. Energy efficiency rating (*sap12*) was chosen as the dependent variable to serve as a proxy indicator for the wide range of features that influence energy use and indoor conditions across the building stock. Its relationship with the independent variables is demonstrated in Equation (1), where the dependent variable and independent variables are on the y-axis and x-axis respectively. Independent variables were first identified by cross-referencing variables from the EHS dataset with those commonly used in previous works. The

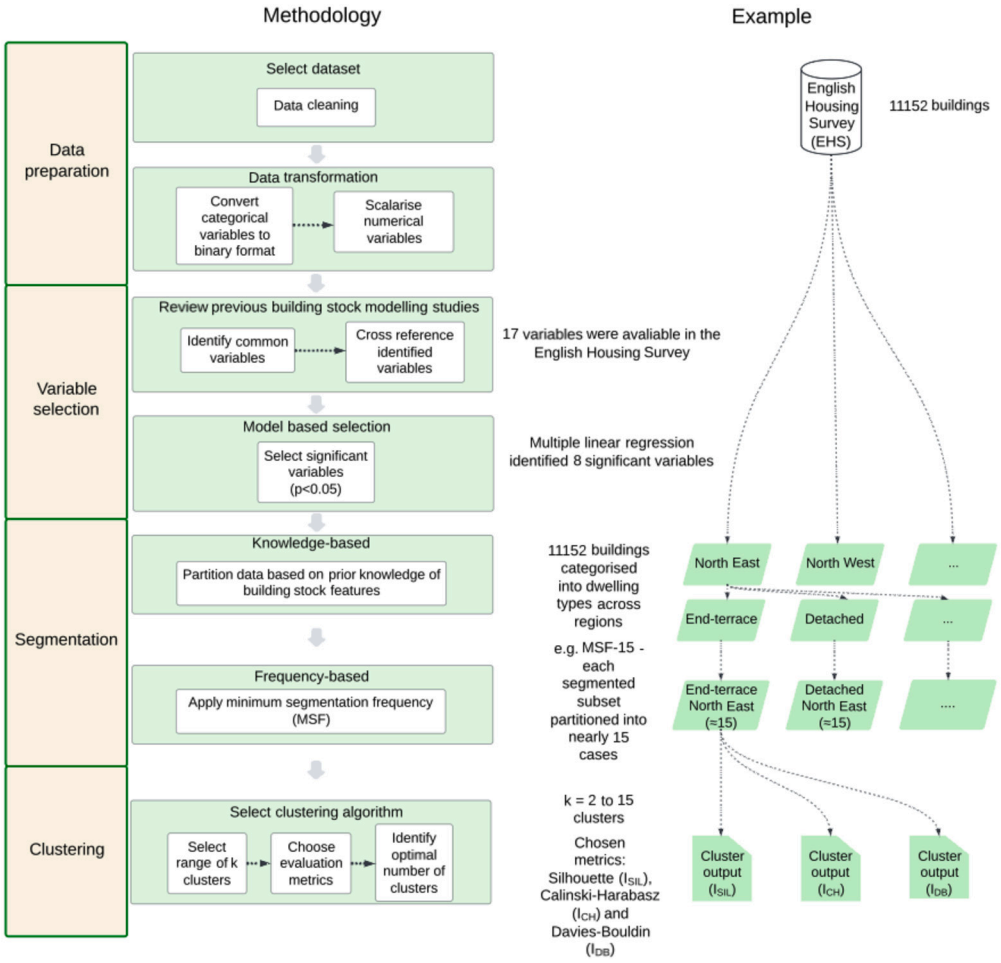


Fig. 1. Overview of the methodology. Example application is given on the right to illustrate the progressive selection of variables and building count.

identified variables and their EHS symbols (in bracket) are: floor area (*floory*), loft insulation thickness (*loftins4*), number of storeys (*storeyx*), boiler system (*boilerx*), fuel type (*fuelx*), system age (*sysage*), dwelling age (*dwage5x*), type of wall and insulation (*wallinsz*), dwelling type (*dwtypenx*), double glazing percentage (*dblglaz2*), heating system (*heat4x*), number of rooms (*nrooms1a*), number of bedrooms (*nbedsx*), income (*hhinc5x*), number of occupants (*hhsizex*), household age (*agehrp2x*) and tenure groups (*tenure2*). The coefficient of determination (R^2) was used to evaluate the regression model, executed using IBM SPSS Statistics (Version: 27.0.1.0). The R^2 value of the regression model was 0.753, predicting roughly three-quarters of the variance in the building stock's energy efficiency ratings. This agrees with the results of earlier research, which found that dwelling geometry, heating system efficiency and wall U-value together account for 75% of the energy efficiency rating [54].

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon \quad (1)$$

where y is the dependent variable, $x_1 \ldots x_n$ are the independent variables β_0 is the constant term when all predictors are zero, $\beta_1 \ldots \beta_n$ are the regression coefficients of the independent variables and ϵ is the residual term.

Household-related variables such as *hhsizex* and *agehrp2x* achieved low regression coefficients, resulting in their exclusion from the final regression model, presented in Table 1. Only *heat4x* and *fuelx* were found to be insignificant, having p-values (Sig.) ≤ 0.05. Hence, it was decided to keep *fuelx* only since retaining it may act as a substitute for both variables *boiler* and *heat4x*. For example, if *fuelx* is gas, the associated *heat4x* will likely be central heating systems, while if *fuelx* is electric, the corresponding *heat4x* would be electrical heating systems. Hence, this approach allows for an optimised variable selection without compromising the representation of the building stock features. In addition, multicollinearity was investigated using variance inflation factors (VIFs) to verify the validity of the regression outputs. An average VIF score of 1.95 suggests a moderate level of multicollinearity between the variables, thereby indicating minimal influence of multicollinearity on the regression outputs [55].

3.3. Segmentation

Pre-clustering segmentation is an important step in ensuring representativeness of the resulting archetypes by avoiding imbalances in the distribution of variables. To achieve this, two segmentation approaches were implemented: knowledge- and frequency-based partitioning.

3.3.1. Knowledge-based

Knowledge-based segmentation groups buildings based on their inherent characteristics, such as dwelling type and region, to account for regional variations. In this study, the EHS data was segmented into 63 distinct subsets based on seven dwelling types (end-terrace, mid-terrace, semi-detached, detached, bungalow, converted flat, and purpose built flat) and nine regions (North East, North West, Yorkshire and the Humber, East Midlands, West Midlands, East of England, London, South East and South West).

3.3.2. Frequency-based

Variables in the segmented data from the previous step often have uneven distributions, with certain features being dominant. This bias

Table 1
Multiple linear regression estimates of influencing variables on the indoor environment.

Independent variables			Unstd. coefficients		Std. coefficients	t-statistics	Sig.
Symbol	Name	Category	β	Std. error	β		
β_0	Constant		67.462	.0342		197.137	.000
floory	Floor area		.023	.001	.119	17.041	< .001
dwtypenx	Dwelling type	End-terrace	1.778	.212	.054	8.404	< .001
		Mid-terrace	6.019	.197	.229	30.495	< .001
		Semi-detached	2.063	.173	.090	11.944	< .001
		Converted flat	9.115	.379	.166	24.046	< .001
		Low rise purpose built flat	8.444	.266	.317	31.794	< .001
		High rise purpose built flat	7.509	2.057	.104	3.651	< .001
dwage5x	Dwelling age	Pre-1919	-9.349	.225	-.366	-41.548	.000
		1919 to 1944	-8.527	.182	-.304	-46.961	.000
		1945 to 1964	-7.480	.152	-.303	-49.081	.000
		1965 to 1980	-5.891	.144	-.245	-40.776	.000
dblglaz2	Double glazing percentage	80% or more double glazed	3.192	.192	.089	16.592	< .001
heat4x	Heating system	Storage heater	1.327	.787	.027	1.687	.092
		Fixed room heater	-9.454	.822	-.136	-11.505	< .001
sysage	System age	Less than 3 years	1.343	.177	.058	7.573	< .001
		More than 12 years	.990	.159	.051	6.225	< .001
fuelx	Type of fuel	Not identified - communal system	1.427	1.245	.020	1.146	.252
		Oil fired system	-7.473	.268	-.147	-27.932	< .001
		Solid fuel	-5.805	.970	-.031	-5.987	< .001
		Electric	-7.952	.926	-.198	-8.587	< .001
boilerx	Boiler type	No boiler	-3.765	1.184	-.107	-3.179	.001
		Standard boiler (floor or wall)	-7.016	.230	-.199	-30.535	< .001
		Back boiler (to fire or stove)	-10.520	.513	-.111	-20.509	< .001
		Combination boiler	-4.030	.284	-.079	-14.166	< .001
		Condensing boiler	-.692	.134	-.029	-5.161	< .001
loftins4	Loft insulation thickness	No roof above	-.617	.229	-.021	-2.697	.007
		None	-11.499	.332	-.178	-34.684	< .001
		Less than 100 m	-3.166	.181	-.090	-17.451	< .001
		100 to 150 mm	-1.673	.123	-.071	-13.607	< .001
storeyx	Number of storeys	1	-.878	.225	-.025	-3.906	< .001
		3	1.701	.153	.063	11.106	< .001
		4	4.011	.315	.070	12.731	< .001
		5	4.679	.509	.047	9.191	< .001
		6	8.285	2.022	.117	4.097	< .001
wallinsz	Type of wall and insulation	Cavity uninsulated	-5.198	.127	-.216	-40.779	.000
		Solid with insulation	1.799	.298	.031	6.040	< .001
		Solid uninsulated	-6.935	.178	-.298	-38.976	.000
		Other	3.973	.397	.050	10.006	< .001

can cause clustering algorithms to overlook less frequent but important features. For instance, given the prevalence of cavity insulated walls in the *wallinsz* variable, the clustering algorithm may only identify archetypes with cavity insulated walls and overlook wall types such as solid walls with insulation. Hence, "minimum segmentation frequency" (MSF) was introduced to retain feature diversity in the segmented data before clustering is applied. The approach divides the segmented data into smaller subsets, each containing a number of cases close to the specified MSF value. For example, MSF-15 represents the division of each of the 63 segmented subsets from the previous step into further subsets, each comprising approximately 15 cases. To examine the influence of MSF on the number and representativeness of the resultant archetypes, a sensitivity analysis was conducted. This involved repeating the frequency-based segmentation step eight times with different MSF values, ranging from 15 to 50, in increments of five.

3.4. Clustering

Clustering is a multivariate classification technique that groups objects into distinct clusters based on their intrinsic characteristics. Objects within the same cluster share comparable characteristics, reflecting a high level of within-cluster coherence while retaining unique distinctions between clusters.

Standard *k*-means and *k*-modes are clustering algorithms for numerical and categorical data respectively. They are not suitable for mixed data because they use different dissimilarity measurements [56]. *k*-means uses the Euclidean distance, which measures the distance between two points in a numerical space. On the other hand, *k*-modes uses the Hamming distance, which is a measure of the difference between two binary vectors. Huang [56] proposed *k*-prototype, which clusters mixed data types using *k*-means' Euclidean distance and *k*-mode's Hamming distance, the first and second expressions in Equation (2) respectively. The algorithm utilises the mean and mode of numerical and categorical variables respectively to minimise dissimilarity between cluster points. Clusters are formed randomly based on the predetermined number of clusters k, the algorithm is then iterated until each cluster's mean and mode values are adjusted and minimised based on the distance between cluster points.

$$d(x,y) = \sum_{i=1}^{p} ||x_i - y_i||^2 + \gamma \sum_{i=p+1}^{q} \delta(x_i, y_i) \tag{2}$$

where the first term represents the Euclidean distance between two numerical datapoints, the second term represents the Hamming distance between two categorical datapoints, and γ and δ are weighting factors to balance numerical and categorical distributions.

3.4.1. Cluster evaluation

The performance and efficacy of clustering techniques are determined using evaluation metrics, which quantify the quality of cluster formations by assessing the cohesiveness of the groupings and how different they are from one another. Clustering evaluation metrics can be divided into two categories: internal and external. Internal metrics measure the quality of the clusters themselves, while external metrics measure the accuracy of the clustering algorithm against known ground

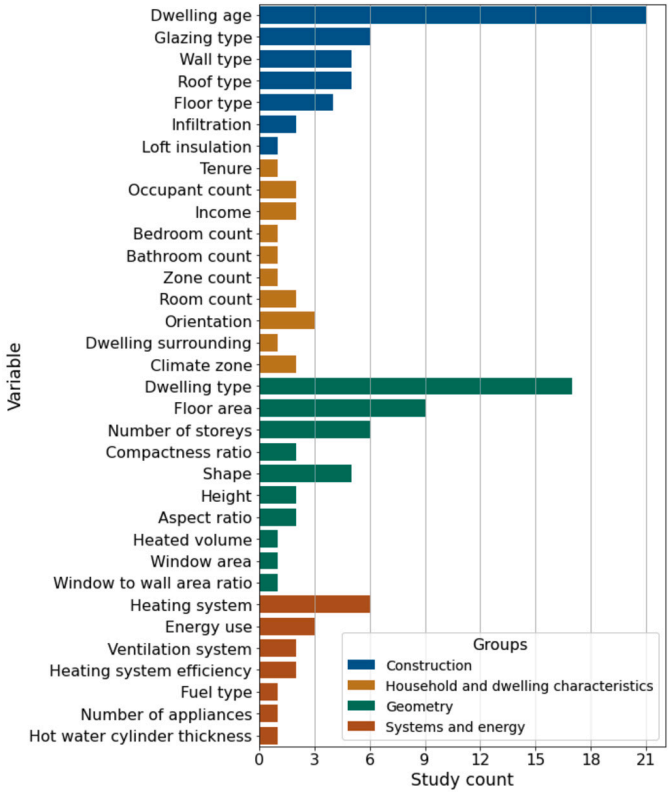


Fig. 2. Variables used in previous building stock modelling works.

truth labels. Internal metrics are commonly used for unsupervised clustering, where ground truth labels are not available. This research investigated the following three most commonly used metrics:

- **Davies-Bouldin index** (I_{DB}) quantifies cluster quality by balancing its compactness and separation, enabling the comparison of solutions and optimisation of cluster numbers [57], as defined in Equation (3). Separation measures the distance between clusters, and compactness measures data point proximity within clusters. Lower values of I_{DB} indicate well-separated, condensed clusters.

$$I_{DB} = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{s_i + s_j}{d_{ij}}\right) \tag{3}$$

  where k is the number of clusters, i and j represent cluster labels where s_i and s_j are cluster samples with respect to their centroids and d_{ij} denotes the distance between the centroids.
- **Silhouette index** (I_{SIL}), also known as Silhouette coefficient, describes the cohesiveness and separation of clusters by comparing the similarity of an object within its cluster to that of the objects in other clusters [58]. Equation (4) is used to calculate I_{SIL}, which ranges from −1 to 1. $I_{SIL} > 0.5$ signifies robust clustering [58] where higher values denote a more distinctive and compact cluster.

$$I_{SIL} = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{4}$$

  where a_i is the average distance between the data point i and all other data points in the same cluster and b_i is the smallest average distance between the data point i and all other data points in the other clusters. Therefore, a_i represents the cohesiveness of the cluster containing the data point i and b_i denotes the extent of separation from the other clusters.
- **Calinski-Harabasz index** (I_{CH}) determines the optimal number of clusters by measuring the separability of clusters, and is calculated using Equations (5) to (7), dividing the total between-cluster dispersion (B_k) by the total within-cluster dispersion (V_k) [59]. A greater value of I_{CH} indicates that the clusters are more distinct from one another and more dense within themselves.

$$B_k = \sum_{i=1}^{k} C_i ||m_i - m||^2 \tag{5}$$

$$W_k = \sum_{i=1}^{k}\sum_{x\in C_i} ||x - m_i||^2 \tag{6}$$

$$I_{CH} = \frac{V_B}{V_W} \times \frac{n-k}{k-1} \tag{7}$$

  where k is the number of clusters, n is the total number of data points, C_i is the size of cluster i, m is the total mean of the dataset, m_i is the mean of cluster i, x is a data point in cluster i, V_B is the average between-cluster sum of squares and V_W is the average within-cluster sum of squares.

3.4.2. Determining the number of clusters

k-prototype clustering algorithm was implemented on the segmented subsets with the value of k ranging from 2 to 15. This range allowed a balanced examination of cluster possibilities while preserving computational feasibility. To determine the number of archetypes for each subset, optimal values of I_{SIL}, I_{CH} and I_{DB} were considered.

3.4.3. Post-processing of clustering outputs

The clustering algorithm assigns each case in the segmented subset to a specific cluster. The algorithm also identifies the centroid of each cluster. Where modelling is relatively straightforward and requires only the variables used in clustering, the centroid can act as the archetype, representing the cluster. In cases where modelling should ideally be based on real cases, the archetype is the closest case from the centroid. The matching of the centroid to a real case allows access to all variables in the original EHS dataset, not just the variables used for clustering. Corresponding dwelling count is then found by aggregating the rounded dwelling weight (*aagpd1920*) values of EHS cases sharing the same cluster number or ID. This step is repeated for all segmented subsets (63 in this research) to identify all archetypes in the EHS dataset.

Modelling the identified representative archetypes in appropriate simulation programs is the next step. Depending on the study objectives, more information than the variables utilised for clustering may be required for modelling individual cases. For example, floor area was used as a clustering variable in this study because of its importance in investigating energy and environmental performance of buildings. However, 3D geometric modelling for energy simulation requires the translation of floor area into building height, width and depth. Instead of making assumptions about the geometry parameters, i.e. width, depth and height, further EHS variables such as ground floor width (*Fdhmwid1*), depth (*Fdhmdep1*) and ceiling height (*cheight0*) can be used to effectively create the 3D geometry of the ground floor of the selected EHS case. Cross-linking the cluster number with the EHS ID (*serialanon*) thus affords the user to extend downstream simulation capabilities in terms of purpose and scope, which is one of the strengths of data-driven archetype identification.

Representative archetype development also offers the benefit of extending the analysis time horizon. For example, future energy and environmental performance under a changing climate can be evaluated using archetypes derived from the current building stock features. Assumptions about the evolution of the building stock such as the changes in heating systems from gas-fired boilers to heat pumps can be encapsulated in multiple scenarios with varying replacement rates, which can then be simulated to investigate the effects of their installation. Assuming that the core features of the current building stock remain unchanged, the representative archetypes can be suitable for assessing how existing buildings might perform under future warming conditions. However, the applicability of the archetypes may be limited in scenar-

ios that involve changes to the core building stock features, i.e. the changes to the variables used for clustering. For instance, if a future scenario considers that a significant share of the new buildings by 2050 will be purpose-built flats with smaller floor area than the present, the characteristics of the building stock will change. In such cases, the baseline archetypes, which are based on the current data, serve as a starting point but may require adaptation or the development of new archetypes to reflect these changes.

3.4.4. Estimating representativeness

The representativeness of the archetypes was determined by comparing the total number of dwellings per variable between the clustering models and EHS using the Mean Absolute Percentage Error (MAPE). The MAPE of the variables was then averaged to indicate the clustering model's overall representativeness, where lower MAPE indicated greater representativeness. The MAPE equation is defined as:

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{8}$$

where n represents the total number of cases, and y_i and $\hat{y}_i$ are the total dwelling count for the variables of the EHS and clustering models respectively.

4. Results and discussion

This section critically discusses the process of clustering for archetype development, using the results from the sensitivity analysis conducted with different minimum segmentation frequencies, clustering evaluation metrics and variable counts. Discussion focuses on representativeness, i.e. the similarity in the distribution of variables between the clustering outputs and EHS dataset.

4.1. Evaluation metric

Different clustering evaluation metrics identified varying number of archetypes, each offering distinct levels of representativeness. Fig. 3 illustrates the number of archetypes and their corresponding representativeness for different clustering evaluation metrics. Across varying levels of MSF, I_{SIL} consistently identified the fewest number of archetypes. This observation may suggest that I_{SIL} tends to identify more uniform clusters, potentially overlooking variations within the building stock, making it less suitable for a comprehensive stock analysis. On the other hand, I_{DB} detected the most archetypes, which can be preferable for studies requiring a thorough representation of building characteristics. While I_{CH} identified fewer archetypes than I_{DB}, it nonetheless demonstrated satisfactory representativeness, attempting to balance the number and representativeness of the archetypes. Thus, while each clustering evaluation metric has its intrinsic strengths and limitations, its strategic selection and application depend on the specific goals and granularity required in the research. By carefully tuning the choice of metric and MSF, researchers can achieve their desirable archetype representativeness, whether they seek a broad overview or a detailed portrayal of the building stock.

4.2. Variable count

The sensitivity analysis also explored the influence of variable count on archetype representativeness. Five variable groupings were investigated, as illustrated in Fig. 4. Across all metrics, a reduction in variable count typically resulted in higher representativeness, suggesting that using fewer clustering variables would result in a smaller number of building archetypes with higher representativeness. I_{SIL} showed the biggest reduction in MAPE with decreasing variable count, followed by I_{CH}, then I_{DB}. The difference in the MAPE between I_{DB} and I_{CH} was considerably smaller than the difference between I_{DB} and I_{SIL}. In addition, I_{CH} identified more archetypes than I_{DB} as the variable count

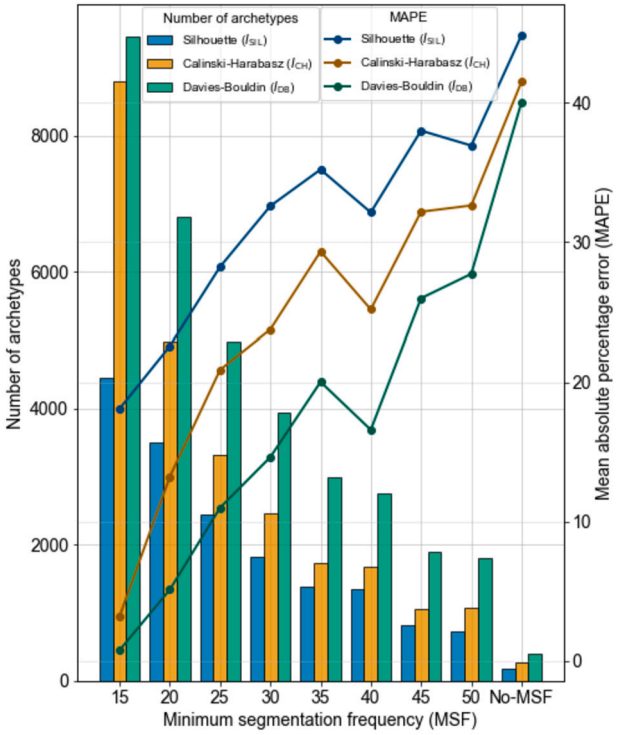


Fig. 3. Representativeness and the number of resulting archetypes with different clustering evaluation metrics and segmentation levels. Lower MAPE on the secondary axis denotes higher representativeness.

decreased. These findings suggest that I_{DB} can be suitable for studies with a variety of variable counts or limited data availability, as it can achieve satisfactory representativeness with a relatively low variable count. Given that I_{DB} demonstrated better performance than I_{SIL} and I_{CH} in identifying archetypes across different variable counts and MSF, it was selected as the primary metric for further investigations. This decision enables a more focused exploration of how similar the distributions of variables in clustered outputs are to that of the EHS data.

4.3. Segmentation level

Segmentation level influences the number of resulting archetypes and their representativeness. Fig. 3 presents the results of the sensitivity analysis of different segmentation levels, illustrating the trade-off between granularity and representativeness. Decreasing the MSF (i.e. increasing the segmentation level) increased the representativeness of the building stock features, but also produced more archetypes, potentially increasing downstream computational costs for simulations. The increased level of representativeness associated with higher segmentation levels is due to the partitioning of the data into incrementally finer subsets, each of which is subjected to clustering. Therefore, users need to carefully select the segmentation level, depending on the granularity and representativeness required for their specific study.

The distribution of variables in the clustering outputs varied considerably for different segmentation levels. Fig. 5 highlights the impact of segmentation level on the distribution of the categorical variables in the clustering outputs. Significant deviations can be observed between the No-MSF and MSF-15 models. The No-MSF model consistently overestimated the share of the dominant features at the expense of less-dominant ones. Hence, the resulting distribution was noticeably different from the distribution in the EHS. On the other hand, the distributions of all seven categorical variables in the MSF-15 outputs were almost similar to that of the distributions in the EHS. The No-MSF model overestimated the categories of *sysage*, *loftins4* and *wallinsz* by 22.5%, 14.5% and 12.5%, respectively. The model overestimated systems aged

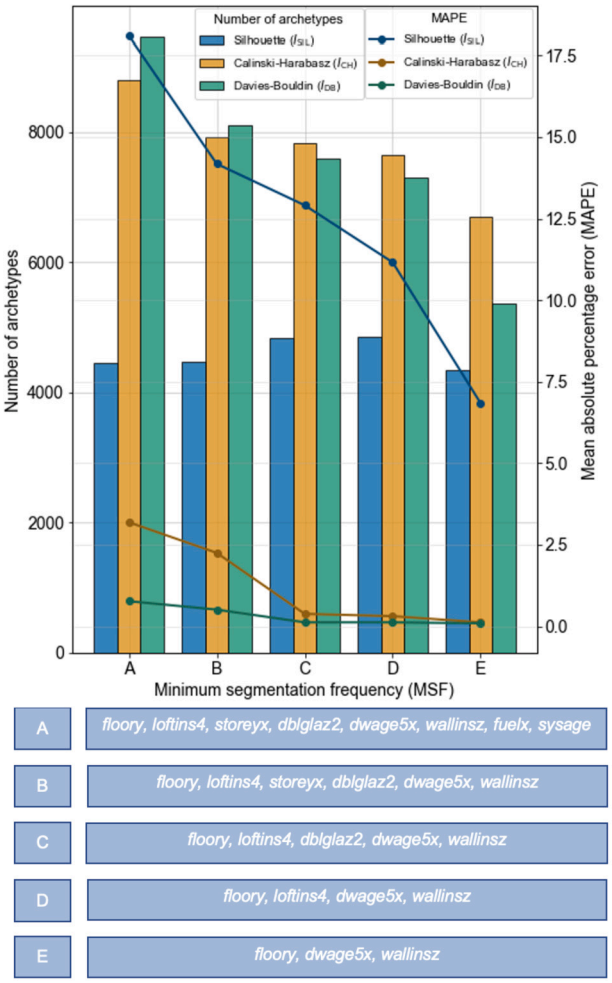


Fig. 4. Representativeness and the number of resulting archetypes with different variable count in clustering. The bar plot represents the number of archetypes while the line plot represents MAPE. A lower MAPE denotes higher representativeness. The analysis consists of the variables *floory* (floor area), *loftins4* (loft insulation thickness), *storeyx* (number of storeys), *dblglaz2* (double glazing percentage), *dwage5x* (dwelling age), *wallinsz* (type of wall and insulation), *fuelx* (type of fuel) and *sysage* (system age).

more than 12 years, buildings with loft insulation thickness of 150 mm or more and cavity insulated buildings. The tendency of the No-MSF model to overestimate building stock characteristics can have significant implications, especially if it is used to inform policy-making or strategic planning. For example, overestimating the prevalence of older systems (as indicated by *sysage*) could suggest that there are more inefficient systems than is the case, which could lead to the misallocation of resources for system replacements. Moreover, misrepresenting the number of buildings with substantial insulation could lead to policymakers believing that buildings are better insulated than they are, delaying essential energy efficiency measures.

The floor area distributions from the clustering outputs are shown in Fig. 6, where the MSF-15 model's distribution was similar to that of the EHS, with an average difference of around 0.66%. In contrast, the No-MSF model's floor area distribution considerably deviated from the EHS. This model particularly underestimated the area of detached houses by approximately 45.5% and, conversely overestimated the area of other dwelling types, with converted flats being the most affected. The No-MSF model's limited ability to accurately represent the building stock's floor area could result in miscalculations of energy demands and efficiency, leading to inadequate or excessive provisions for heating, cooling and lighting. Within the EHS, detached dwellings showed significant variance in floor area distribution, potentially causing the clustering algorithm to focus on the most common sizes, overlooking larger dwellings. Furthermore, the limited sample size of converted flats may have constrained the algorithm's ability to effectively learn, possibly increasing its sensitivity to anomalies and skewing the overall representation. Therefore, adopting frequency-based segmentation, e.g. MSF-15, is essential to mitigate these discrepancies observed in the No-MSF model, and to provide a more accurate representation of the building stock's floor area distribution.

The MSF-15 model's ability to represent the building stock is further demonstrated by its close alignment with the EHS's distribution of dwelling types across different regions, as shown in Fig. 7. For example, in London, the MSF-15 model's distribution of end- and mid-terrace dwellings differed from the EHS by less than 0.1%, while its distribution of purpose-built flats differed by approximately 0.43%. In contrast, the No-MSF model underrepresented London flats by 3.94% and incorrectly identified the South West region as having the most purpose-built flats. Misidentifying regions with predominant dwelling types can skew regional development plans, potentially causing overcrowding or under-utilisation, which may lead to ineffective housing and urban development strategies.

5. Archetype development framework

Insights gained from the sensitivity analysis informed the development of a comprehensive framework for guiding the creation of building archetypes. The framework allows the user to consider the interaction between influencing and decision factors during the archetype development process. As presented in Table 2, the framework comprises four influencing factors: geographical scale, research focus, temporal scale and computational cost. Given a set of influencing factors pertinent to the specific archetype development study, a user can choose the corresponding recommended values of the three decision factors: minimum segmentation frequency (MSF), evaluation metric and variable count.

Influencing factors are broadly categorised into features. The geographical scale is divided into district, city and national, based on stock homogeneity. Research focus is classed into specific and broad, depending on how focused the study objectives are. The specific research focus is linked with the investigation into specific characteristics of the building stock, typically within a single domain, e.g. energy efficiency. Whereas the broad research focus is typically multi-domain and requires the modelling of interdependent factors, e.g. energy and environmental performance and the cost of retrofitting. Another way to differentiate between 'specific' and 'broad' research focus is to look at the number of dependent variables needed to identify significant variables for use in clustering using regression analysis. Specific research would normally require one dependent variable, whereas the broad focus might involve multivariate regression analysis with two or more dependent variables. Temporal scales range between short- and long-term, referring to instantaneous to monthly and annual to decadal respectively. Computational cost depends on the detail and number of domains being modelled. Hence it is characterised by two features: low and high, with the assumption being that simplified or steady-state models are computationally less expensive than detailed and dynamic models to simulate the building archetypes.

MSF is inversely linked with segmentation level, i.e. the number of resulting data partitions from frequency-based segmentation. In this framework, MSF is divided into low, moderate and high with corresponding values of less than 25, between 25 and 40, and more than 40 respectively. There are three clustering evaluation metrics in the framework: Calinski-Harabasz (I_{CH}), Davies-Bouldin (I_{DB}) and Silhouette (I_{SIL}). Variable count refers to the number of variables selected for clustering. Even though the regression results may suggest a higher number of significant variables within the building stock data, the user may opt to use fewer variables to handle multicollinearity and reduce computation time during clustering. In the framework, low, moderate

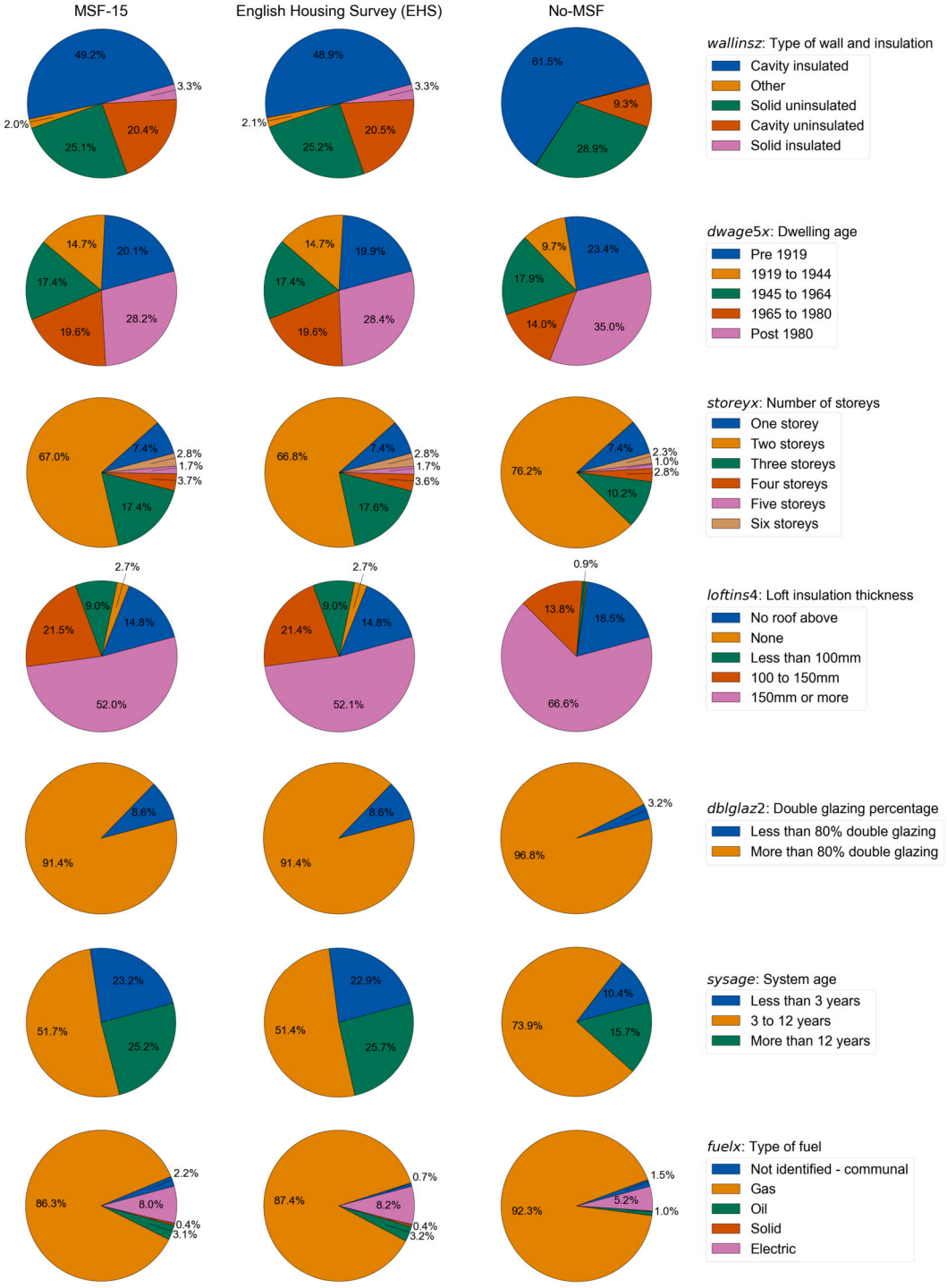


Fig. 5. Comparison of the distribution of categorical variables in the clustering outputs and EHS.

and high variable count refers to between 2 and 3, between 4 and 6, and more than 6 variables respectively.

5.1. Geographical scale

The methodological approach to archetype development is significantly influenced by the geographical context [31]. Neighbourhoods and homogeneous districts are often characterised by limited data availability [60]. In such cases, simplified models with few variables are generally more applicable than detailed models that require disaggregated data. A low segmentation level, i.e. high MSF, is often sufficient for neighbourhoods and districts due to the homogeneity in building characteristics such as age, materials, construction and usage. This reduced complexity avoids the unnecessary partitioning of data, as the buildings are likely homogeneous enough to be adequately represented with fewer archetypes. I_{SIL} can be ideal in these circumstances, as demonstrated in Fig. 3, as the index consistently identified the fewest archetypes. However, if increased representativeness is desired within the scope of the available computational resources, I_{DB} can be employed to provide a more comprehensive portrayal of the building stock.

Conversely, the likelihood of the existence of high-quality data is higher for urban and national contexts, which supports the use of more

Table 2
Framework for developing building archetypes considering different influencing factors.

Influencing factor	Feature	MSF[1]	Evaluation metric[2]	Variable count[3]
Geographical scale	District	High	I_{SIL} or I_{DB}	Low
	City	Low to moderate	I_{DB} or I_{CH}	Moderate to high
	National	Low	I_{DB} or I_{CH}	Moderate to high
Research focus	Specific	Moderate to high	I_{SIL} or I_{DB}	Low
	Broad	Moderate to high	I_{DB} or I_{CH}	Low to moderate
Temporal scale	Short-term	Low	I_{DB} or I_{CH}	Moderate to high
	Long-term	Moderate to high	I_{DB}	Low
Computational cost	Low (e.g. steady-state simulation)	Low	I_{DB} or I_{CH}	Moderate to high
	High (e.g. dynamic simulation)	Low to moderate	I_{DB}	Low to moderate

[1] Minimum segmentation frequency (MSF): Low (MSF: < 25), Moderate (MSF: 25-40), High (MSF: > 40).
[2] Evaluation metric: I_{CH} (Calinski-Harabasz), I_{DB} (Davies-Bouldin), I_{SIL} (Silhouette).
[3] Variable count: Low (2-3), Moderate (4-6), High (> 6).

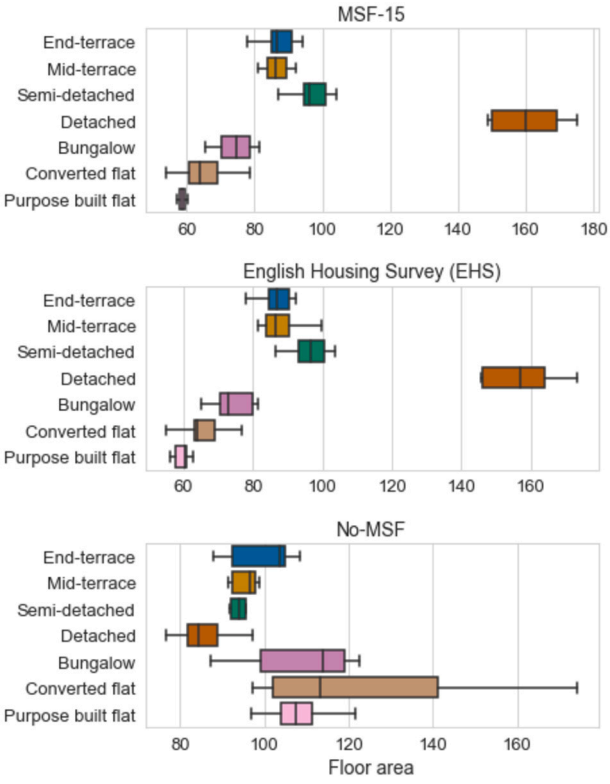


Fig. 6. Floor area distribution of the clustering outputs and EHS.

complex modelling. Higher levels of segmentation, i.e. low MSF, can be adopted in such cases. In heterogeneous larger geographies, I_{CH} and I_{DB} indices are more applicable as they are better suited in identifying a broader range of representative archetypes, as shown in Fig. 3. Representativeness is crucial in large-scale studies for accurately representing the variety of building characteristics found in heterogeneous building stocks, to ensure a detailed and encompassing view of the urban and national building landscapes.

5.2. Research focus

Research focus in building stock modelling varies from specific studies targeting a single domain to broader analyses considering multiple domains. The particular needs of the study, regardless of the geographical scale, often leads to the adoption of varying number of archetypes. For instance, a specific research on the effects of increasing cavity wall insulation on internal temperatures, a low to moderate segmentation level is often sufficient, particularly as the dataset tends to be uniform in insulation characteristics. The homogeneity of the data in this case facilitates the adoption of fewer variables. I_{SIL} is an ideal choice for evaluation metric because it identifies the fewest archetypes, as shown in Fig. 3. However, for studies that target all types of wall insulation, I_{DB} may be more appropriate, as it can handle more variables and is capable of identifying archetypes with high representativeness.

Modelling complexity increases in studies with a broader research focus and multiple objectives. For instance, studies on indoor overheating due to climate change and corresponding energy demand require the consideration of complex interactions between two interconnected domains: building thermal dynamics and energy systems. To effectively address this dual focus, the study would likely require a multivariate regression analysis to identify relevant clustering variables, using at least two dependent variables: indoor temperature and energy consumption. Hence, higher segmentation levels and more variables may be needed to adequately model the variations in building thermal characteristics, and energy and environmental systems to study their influence on indoor temperature and energy demand. When high segmentation levels are required, the choice between I_{CH} and I_{DB} can be guided by the variable count and availability of computational resources. I_{CH} appeared to be more effective for high variable counts, as shown in Fig. 4, as it identifies fewer archetypes than I_{DB}, albeit at the expense of representativeness. However, in cases where computational cost is not a concern, I_{DB} can be a better choice for enhanced representativeness.

5.3. Temporal consideration

Building stock modelling studies focusing on short-term analysis may require archetypes that comprehensively represent existing building characteristics. Hence, representative archetypes are essential for ensuring relevant analyses to inform effective decision-making and policy formulation. Figs. 3-7 demonstrate how higher segmentation levels are well-suited to achieving high representativeness of variables such as “type of wall and insulation” (*wallinsz*), and are capable of adequately capturing different building typologies with a floor area variation closely resembling that of the original building stock, i.e. the EHS data. Among all clustering evaluation metrics, I_{DB} is found to be particularly effective for such comprehensive analyses as it achieved the highest representativeness with a low variable count, as shown in Fig. 4.

For long-term building stock analyses that anticipate changes in building characteristics, a low to moderate level of segmentation, and a low variable count are recommended to minimise computational costs and avoid the risk of archetypes becoming inconsistent or irrelevant over time. For instance, referring to Fig. 5, it is observed that the No-MSF (i.e. no frequency-based segmentation) model tends to overestimate the prevalence of cavity-insulated dwellings within the existing building stock. However, this overestimation might be considered less-critical when projecting future (e.g. by 2050 or 2100) scenarios for indoor overheating assessment, given the anticipated rise in newly constructed dwellings featuring cavity wall insulation.

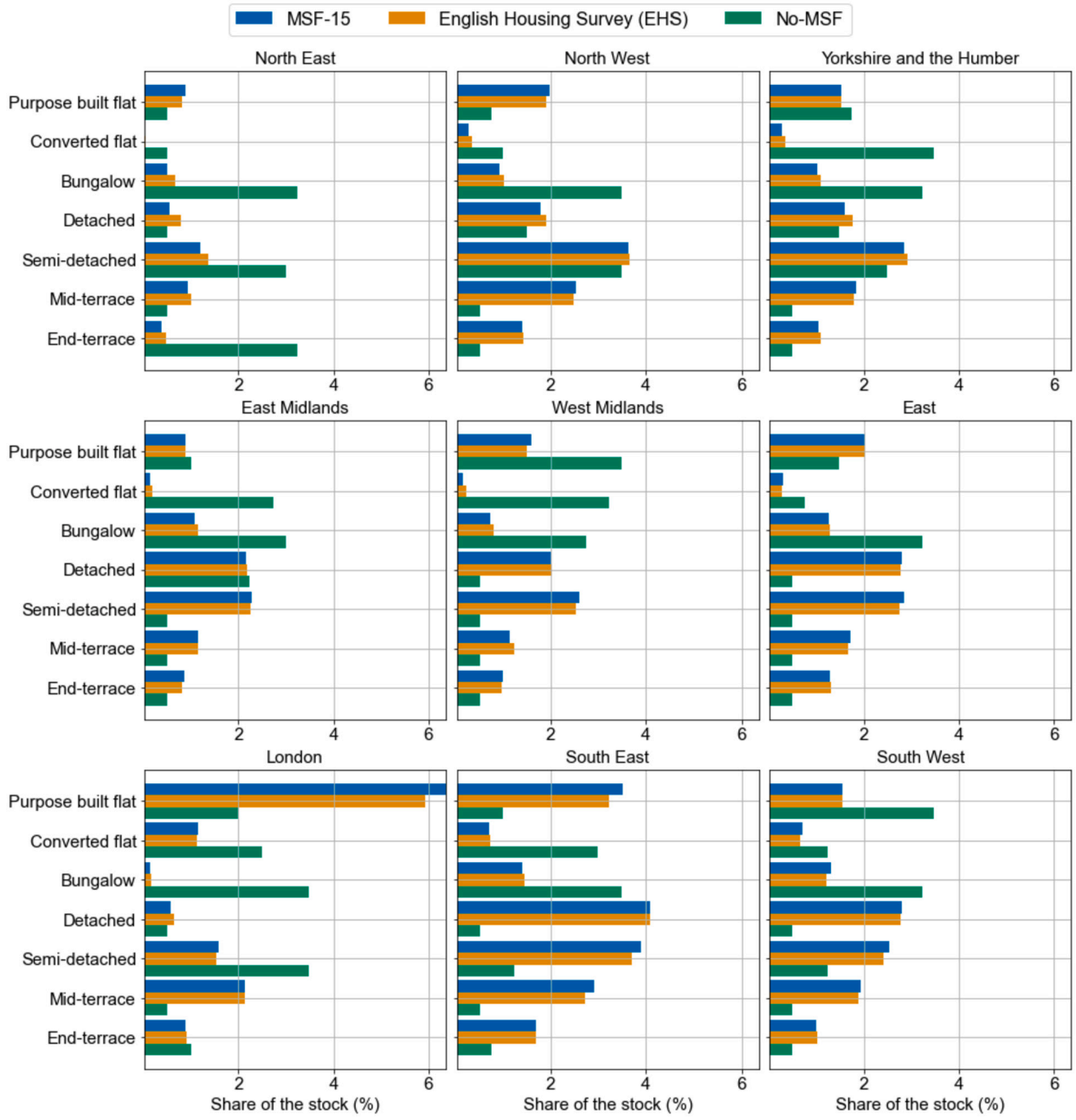


Fig. 7. The distribution of dwelling types and regions between the clustering outputs and EHS.

5.4. Computational cost

Simplified modelling such as steady-state simulations, due to their relatively lower computational demands [61], are well-suited for building stock studies comprising a wide range of archetypes. Simplified models are often able to deal with diverse variables from multiple domains. A high level of segmentation is, therefore, recommended to account for the diversity of variables. I_{DB} is typically preferred in these scenarios for its ability to handle a variety of clustering variables, as shown in Fig. 4. However, I_{CH} can also be used, especially when fewer archetypes are sufficient, offering flexibility in archetype development.

In contrast, detailed modelling such as whole-building dynamic simulations are computationally expensive [62], requiring careful considerations of the impact of the selected segmentation level on the number of resulting archetypes. While a larger number of archetypes can fully leverage the capabilities of dynamic simulations to provide detailed temporal insights, researchers often face limitations in computational power. This consideration becomes particularly crucial as the geographical scope increases and the building stock becomes more diverse. When less resources, both time and computation, are available to the user, adopting a moderate to low segmentation level can be a practical approach to acquiring meaningful insights while dealing with resource constraints. This approach limits the dataset from being overly segmented, thereby reducing the number of archetypes needed for individual dynamic simulations. The combination of I_{DB} and low variable count can be advantageous for representative clustering, as demonstrated in Fig. 4, especially when dynamic simulations are used as the analysis tool.

6. Limitations and future works

The use of MSF improved archetype representativeness noticeably. However, its effectiveness is sensitive to the distribution of the variables, being less pronounced for skewed distributions. Future research

can explore alternative segmentation approaches to account for the skewness in the data. Clustering can also be investigated without setting thresholds on the number of cases per segmented subset. Further research could also look into the application of MSF in conjunction with clustering algorithms other than *k*-prototype. These investigations may enable the exploration of further objectives, including minimising the number of archetypes while maintaining sufficient representativeness, even with increasing variable counts.

Compared with the larger geographies such as the United States and China, English housing stock can be considered homogeneous, despite noticeable regional variations. On the other hand, dense cities in the developing Asia are often characterised by a larger share of multifamily buildings that are more homogeneous in nature than the English housing stock. The generalisability of the proposed approach for the development of representative archetypes can be investigated in other contexts of varying homogeneity and stock characteristics.

The physical, thermal and system characteristics of non-domestic buildings vary significantly depending on building type, use and location. Although the use of MSF for pre-clustering segmentation resulted in higher representativeness for the investigated dwelling stock, further research should be conducted on how well the combined MSF and *k*-prototype work on non-domestic building stock, particularly focusing on the effects of knowledge- and frequency-based segmentation on representativeness.

7. Conclusion

This study addressed an important research gap in the building stock literature by investigating the effects of segmentation level, clustering evaluation metric and variable count on archetype representativeness. The research introduced the concept of "minimum segmentation frequency" or MSF as a pre-clustering partitioning strategy to improve archetype representativeness. Based on a review of the suitability and effectiveness of clustering algorithms used in past research on building stock modelling, the *k*-prototype algorithm was chosen as the most appropriate method. The widely studied 2020 English Housing Survey (EHS) was chosen as stock data, which was partitioned using MSF before applying the *k*-prototype algorithm on each segmented subset. The findings have important implications for the development of simulation-based building stock and urban modelling, where achieving representativeness with the fewest possible archetypes is a primary objective.

The choice of segmentation level, clustering evaluation metric and variable count influenced the representativeness of the archetypes. Increased segmentation levels led to more archetypes generated through clustering, often with higher representativeness at each level. The Davies-Bouldin index consistently identified the most archetypes and achieved the highest representativeness, followed by the Calinski-Harabasz and Silhouette indices. When a low variable count was adopted, all the indices typically identified fewer archetypes with higher representativeness than when higher variable counts were used for clustering.

The insights gained through the sensitivity analysis facilitated the development of a comprehensive framework for generating representative archetypes. The framework considers factors influencing the development of building archetypes such as geographical and temporal scales, computational cost and research focus. Researchers working on representative archetype development may find the following recommendations useful, for selecting the segmentation level, clustering evaluation metric and variable count:

- Lower segmentation levels can be suitable for district-scale studies with homogeneous building stock and when more resource-intensive dynamic simulations are needed. Whereas higher segmentation levels are better suited for more heterogeneous city- and national-level stocks, and when steady-state simulations are sufficient.
- If computational resources are not a limiting factor, the Davies-Bouldin index can be an effective metric for achieving high archetype representativeness. For resource-limited scenarios, the Calinski-Harabasz index offers a viable alternative, achieving a balance between representativeness and computational cost by identifying fewer archetypes. However, the Calinski-Harabasz index may not be ideal for clustering with few variables. The Silhouette index can be suitable for building stocks with one or more dominant variables, or for studies with specific objectives, as it consistently identified the least number of archetypes in this study.
- For national-scale studies with specific objectives that require dynamic simulations over a long-horizon, reducing the number of variables used for clustering can be beneficial. This approach simplifies the complexities of national landscapes, reduces computational cost, and avoids producing overly detailed archetypes that may become less relevant in the future as building trends and technologies evolve.

CRediT authorship contribution statement

Mousa Alrasheed: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Monjur Mourshed:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] M. Röck, E. Baldereschi, E. Verellen, A. Passer, S. Sala, K. Allacker, Environmental modelling of building stocks – an integrated review of life cycle-based assessment models to support EU policy making, Renew. Sustain. Energy Rev. 151 (2021) 111550.
[2] J.L. Reyna, M.V. Chester, Energy efficiency to reduce residential electricity and natural gas use under climate change, Nat. Commun. 8 (2017) 14916.
[3] D. Wang, J. Landolt, G. Mavromatidis, K. Orehounig, J. Carmeliet, CESAR: a bottom-up building stock modelling tool for Switzerland to address sustainable energy transformation strategies, Energy Build. 169 (2018) 9–26.
[4] Y. Yamaguchi, B. Kim, T. Kitamura, K. Akizawa, H. Chen, Y. Shimoda, Building stock energy modeling considering building system composition and long-term change for climate change mitigation of commercial building stocks, Appl. Energy 306 (2022) 117907.
[5] J. Pittam, P.D. O'Sullivan, G. O'Sullivan, Stock aggregation model and virtual archetype for large scale retrofit modelling of local authority housing in Ireland, Energy Proc. 62 (2014) 704–713.
[6] A. Stephan, A. Athanassiadis, Quantifying and mapping embodied environmental requirements of urban building stocks, Build. Environ. 114 (2017) 187–202.
[7] M. Alrasheed, M. Mourshed, Domestic overheating risks and mitigation strategies: the state-of-the-art and directions for future research, Indoor Built Environ. 32 (2023) 1057–1077.
[8] R. Gupta, M. Gregg, Preventing the overheating of English suburban homes in a warming climate, Build. Res. Inf. 41 (2013) 281–300.
[9] M. Gangolells, M. Casals, Resilience to increasing temperatures: residential building stock adaptation through codes and standards, Build. Res. Inf. 40 (2012) 645–664.
[10] J. Taylor, C. Shrubsole, M. Davies, P. Biddulph, P. Das, I. Hamilton, S. Vardoulakis, A. Mavrogianni, B. Jones, E. Oikonomou, The modifying effect of the building envelope on population exposure to PM2.5 from outdoor sources, Indoor Air 24 (2014) 639–651.
[11] A. Nutkiewicz, R.K. Jain, Exploring the integration of simulation and deep learning models for urban building energy modelling and retrofit analysis, in: Building Simulation Conference Proceedings, 2019, pp. 3209–3216.

[12] A. Mastrucci, O. Baume, F. Stazi, U. Leopold, Estimating energy savings for the residential building stock of an entire city: a GIS-based statistical downscaling approach applied to rotterdam, Energy Build. 75 (2014) 358–367.
[13] K.N. Streicher, P. Padey, D. Parra, M.C. Bürer, S. Schneider, M.K. Patel, Analysis of space heating demand in the Swiss residential building stock: element-based bottom-up model of archetype buildings, Energy Build. 184 (2019) 300–322.
[14] H. Wang, Z.J. Zhai, Advances in building simulation and computational techniques: a review between 1987 and 2014, Energy Build. 128 (2016) 319–335.
[15] S. Hu, D. Yan, E. Azar, F. Guo, A systematic review of occupant behavior in building energy policy, Build. Environ. 175 (2020) 106807.
[16] M. Shahrestani, R. Yao, G.K. Cook, A review of existing building benchmarks and the development of a set of reference office buildings for England and Wales, Intell. Build. Int. 6 (2014) 41–64.
[17] C. Cerezo Davila, C.F. Reinhart, J.L. Bemis, Modeling Boston: a workflow for the efficient generation and maintenance of urban building energy models from existing geospatial datasets, Energy 117 (2016) 237–250.
[18] A.A. Famuyibo, A. Duffy, P. Strachan, Developing archetypes for domestic dwellings - an Irish case study, Energy Build. 50 (2012) 150–157.
[19] J. Dong, Y. Schwartz, A. Mavrogianni, I. Korolija, D. Mumovic, A review of approaches and applications in building stock energy and indoor environment modelling, Build. Serv. Eng. Res. Technol. 44 (2023) 333–354.
[20] Department for Levelling Up, Housing and Communities, Ministry of Housing, Communities and Local Government, English housing survey, 2020: Housing stock data: Special licence access, UK Data Service, 2023, Accession Number: SN 9076.
[21] J. Taylor, P. Wilkinson, M. Davies, B. Armstrong, Z. Chalabi, A. Mavrogianni, P. Symonds, E. Oikonomou, S.I. Bohnenstengel, Mapping the effects of urban heat island, housing, and age on excess heat-related mortality in London, Urban Clim. 14 (2015) 517–528.
[22] M. Rajput, G. Augenbroe, B. Stone, M. Georgescu, A. Broadbent, S. Krayenhoff, E. Mallen, Heat exposure during a power outage: a simulation study of residences across the metro Phoenix area, Energy Build. 259 (2022) 111605.
[23] T. Dalla Mora, L. Teso, L. Carnieletto, A. Zarrella, P. Romagnoni, Comparative analysis between dynamic and quasi-steady-state methods at an urban scale on a social-housing district in Venice, Energies 14 (2021) 5164.
[24] A. Mavrogianni, P. Wilkinson, M. Davies, P. Biddulph, E. Oikonomou, Building characteristics as determinants of propensity to high indoor summer temperatures in London dwellings, Build. Environ. 55 (2012) 117–130.
[25] J. Taylor, M. Davies, A. Mavrogianni, Z. Chalabi, P. Biddulph, E. Oikonomou, P. Das, B. Jones, The relative importance of input weather data for indoor overheating risk assessment in dwellings, Build. Environ. 76 (2014) 81–91.
[26] R. Gupta, M. Gregg, Using UK climate change projections to adapt existing English homes for a warming climate, Build. Environ. 55 (2012) 20–42.
[27] S. Gendebien, E. Georges, S. Bertagnolio, V. Lemort, Methodology to characterize a residential building stock using a bottom-up approach: a case study applied to Belgium, Int. J. Sustain. Energy Plan. Manag. 4 (2014) 71–87.
[28] L.D. Shorrock, J.E. Dunster, The physically-based model BREHOMES and its use in deriving scenarios for the energy use and carbon dioxide emissions of the UK housing stock, Energy Policy 25 (1997) 1027–1037.
[29] L. Pistore, G. Pernigotto, F. Cappelletti, P. Romagnoni, A. Gasparella, From energy signature to cluster analysis: comparison between different clustering algorithms, in: Building Simulation Conference Proceedings, 2017, pp. 469–477.
[30] I. De Jaeger, G. Reynders, C. Callebaut, D. Saelens, A building clustering approach for urban energy simulations, Energy Build. 208 (2020) 109671.
[31] U. Ali, M.H. Shamsi, C. Hoare, E. Mangina, J. O'Donnell, A data-driven approach for multi-scale building archetypes development, Energy Build. 202 (2019) 109364.
[32] G. Tardioli, R. Kerrigan, M. Oates, J. O'Donnell, D.P. Finn, Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach, Build. Environ. 140 (2018) 90–106.
[33] X. Li, R. Yao, M. Liu, V. Costanzo, W. Yu, W. Wang, A. Short, B. Li, Developing urban residential reference buildings using clustering analysis of satellite images, Energy Build. 169 (2018) 417–429.
[34] P. Borges, O. Travesset-Baro, A. Pages-Ramon, Hybrid approach to representative building archetypes development for urban models – a case study in Andorra, Build. Environ. 215 (2022) 108958.
[35] E.L. Ofetotse, E.A. Essah, R. Yao, Evaluating the determinants of household electricity consumption using cluster analysis, J. Build. Eng. 43 (2021) 102487.
[36] K. Echlouchi, M. Ouardouz, A.S. Bernoussi, Eco-district, an ideal framework to initiate large-scale urban energy renovation in Morocco, J. Ecol. Eng. 23 (2022) 100–114.
[37] P. Murray, J. Marquant, M. Niffeler, G. Mavromatidis, K. Orehounig, Optimal transformation strategies for buildings, neighbourhoods and districts to reach CO_2 emission reduction targets, Energy Build. 207 (2020) 109569.
[38] M.M. Madbouly, S.M. Darwish, N.A. Bagi, M.A. Osman, Clustering big data based on distributed fuzzy k-medoids: an application to geospatial informatics, IEEE Access 10 (2022) 20926–20936.
[39] G. Preud'homme, K. Duarte, K. Dalleau, C. Lacomblez, E. Bresso, M. Smaïl-Tabbone, M. Couceiro, M.D. Devignes, M. Kobayashi, O. Huttin, J.P. Ferreira, F. Zannad, P. Rossignol, N. Girerd, Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark, Sci. Rep. 11 (2021) 4202.
[40] S. Lechtenböhmer, A. Schüring, The potential for large-scale savings from insulating residential buildings in the EU, Energy Effic. 4 (2011) 257–270.
[41] J.M.R. Portella, Bottom-up description of the French building stock, including archetype buildings and energy demand, MSc thesis, Chalmers University of Technology, Göteborg, Sweden, 2012.
[42] C. Molina, M. Kent, I. Hall, B. Jones, A data analysis of the Chilean housing stock and the development of modelling archetypes, Energy Build. 206 (2020) 109568.
[43] E. Mata, A.S. Kalagasidis, F. Johnsson, A modelling strategy for energy, carbon, and cost assessments of building stocks, Energy Build. 56 (2013) 100–108.
[44] S.K. Firth, K.J. Lomas, A.J. Wright, Targeting household energy-efficiency measures using sensitivity analysis, Build. Res. Inf. 38 (2010) 25–41.
[45] C. Loucari, J. Taylor, R. Raslan, E. Oikonomou, A. Mavrogianni, Retrofit solutions for solid wall dwellings in England: the impact of uncertainty upon the energy performance gap, Build. Serv. Eng. Res. Technol. 37 (2016) 614–634.
[46] I. Ballarini, V. Corrado, A new methodology for assessing the energy consumption of building stocks, Energies 10 (2017) 1102.
[47] M.S. Geraldi, E. Ghisi, Data-driven framework towards realistic bottom-up energy benchmarking using an artificial neural network, Appl. Energy 306 (2022) 117960.
[48] DLUHC, 50 years of the English Housing Survey, Department for Levelling Up, Housing and Communities (DLUHC), London, UK, 2017.
[49] D.A. Sass, T.A. Schmitt, H.W. Marsh, Evaluating model fit with ordered categorical data within a measurement invariance framework: a comparison of estimators, structural equation modeling: a, Multidiscipl. J. 21 (2014) 167–180.
[50] C.H. Li, Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares, Behav. Res. Methods 48 (2016) 936–949.
[51] D. Olson, D. Delen, Advanced Data Mining Techniques, Springer, 2008.
[52] J.A. Paravantis, M. Santamouris, An analysis of indoor temperature measurements in low- and very-low-income housing in Athens, Greece, Adv. Build. Energy Res. 10 (2016) 20–45.
[53] S.-J. Cao, Challenges of using cfd simulation for the design and online control of ventilation systems, Indoor Built Environ. 28 (2019) 3–6.
[54] A. Stone, D. Shipworth, P. Biddulph, T. Oreszczyn, Key factors determining the energy rating of existing English houses, Build. Res. Inf. 42 (2014) 725–738.
[55] N. Shrestha, Detecting multicollinearity in regression analysis, Am. J. Appl. Math. Stat. 8 (2020) 39–42.
[56] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Min. Knowl. Discov. 12 (1998) 283–304.
[57] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, Understanding of internal clustering validation measures, in: IEEE International Conference on Data Mining, 2010, pp. 911–916.
[58] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[59] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. 3 (1974) 1–27.
[60] M. Turrin, B. Peters, W. O'Brien, R. Stouffs, T. Dogan (Eds.), Proceedings of the Symposium on Simulation for Architecture and Urban Design 2017: SimAUD 2017, Simulation Councils, 2017.
[61] D. Gatt, C. Yousif, M. Cellura, L. Camilleri, F. Guarino, Assessment of building energy modelling studies to meet the requirements of the new energy performance of buildings directive, Renew. Sustain. Energy Rev. 127 (2020) 109886.
[62] T. Hong, Y. Chen, X. Luo, N. Luo, S.H. Lee, Ten questions on urban building energy modeling, Build. Environ. 168 (2020) 106508.