



Title	A data-driven approach for multi-scale building archetypes development
Authors(s)	Ali, Usman, Shamsi, Mohammad Haris, Hoare, Cathal, Mangina, Eleni, O'Donnell, James
Publication date	2019-11
Publication information	Ali, Usman, Mohammad Haris Shamsi, Cathal Hoare, Eleni Mangina, and James O'Donnell. "A Data-Driven Approach for Multi-Scale Building Archetypes Development." Elsevier, November 2019. https://doi.org/10.1016/j.enbuild.2019.109364 .
Publisher	Elsevier
Item record/more information	http://hdl.handle.net/10197/10998
Publisher's statement	This is the author's version of a work that was accepted for publication in Energy and Buildings. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Energy and Buildings (202, (2019)) https://doi.org/10.1016/j.enbuild.2019.109364
Publisher's version (DOI)	10.1016/j.enbuild.2019.109364

Downloaded 2026-05-01 23:37:38

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)



© Some rights reserved. For more information

A data-driven approach for multi-scale building archetypes development

Usman Ali^a, Mohammad Haris Shamsi^a, Cathal Hoare^a, Eleni Mangina^b, James O'Donnell^a

^a*School of Mechanical and Materials Engineering and UCD Energy Institute, UCD, Dublin, Ireland*

^b*School of Computer Science and UCD Energy Institute, UCD, Dublin, Ireland*

Abstract

Globally the building sector accounts for a significant portion of the overall energy demand and greenhouse gas emissions of any country. The most common approach for the collection of modeling and benchmarking data that can be used for predictions of energy performance at a national or urban scale is through classification of the building stock into representative archetypes. Developing such building archetypes is a complex task due to the difficulties associated with gathering detailed geometric and non-geometric data at an urban scale. Although existing databases and projects provide a valuable overview of a building stock, the information about buildings' physical descriptions are not regularly updated. Moreover, these databases cover only the national top-level archetypes and lack crucial information related to city or district scale building stocks. The use of national scale archetypes requires many assumptions that may not hold true for energy modeling at urban or district scale.

This paper proposes a multi-scale (national, city, county and district) archetype development methodology using different data-driven approaches. The methodology consists of following five steps: 1) data collection, 2) segmentation, 3) characterization, 4) quantification, and 5) modeling results. We developed a test case based on the available building stock data of Ireland. The test case used previously developed archetype geometries coupled with the parameters determined by the characterization process to calculate annual energy use (kWh) of buildings at a multiple-scales. The resulting archetypes at national, city, county and district scale are analyzed and compared against one another. The results indicate that significant differences occur in terms of energy modeling results when national scale archetypes are used to simulate the energy performance of buildings at the local scale. These multi-scale building archetypes will aid local authorities and city planners when analyzing energy efficiency and consequently, help to improve sustainable energy policy decisions.

Keywords: building archetypes, building energy efficiency, urban building energy modeling, building energy performance simulation

Email addresses: usman.ali@ucdconnect.ie (Usman Ali), mohammad.shamsi@ucdconnect.ie (Mohammad Haris Shamsi), cathal.hoare@ucd.ie (Cathal Hoare), eleni.mangina@ucd.ie (Eleni Mangina), james.odonnell@ucd.ie (James O'Donnell)

Preprint submitted to Energy and Buildings

August 15, 2019

Nomenclature

\bar{x}	mean
AD	Average within Distance
$ASIEPI$	ASsessment and Improvement of the EPBD Impact
$BEPS$	Building Energy Performance Simulation
$BPIE$	Buildings Performance Institute Europe
BRE	Building Research Establishment
C_i	i^{th} cluster
c_i	Centroid for cluster i
C_j	j^{th} cluster
CRB	Commercial Reference Buildings
CSO	Central Statistics Office
D_{ij}	Separation between the i^{th} and the j^{th} clusters
DBI	Davies Bouldin Index
$DEAP$	Dwelling Energy Assessment Procedure
DOE	Department of Energy
$ENTRANZE$	ENforce the TRANsition to Nearly Zero Energy buildings
$EPBD$	European Union Energy Performance of Buildings Directive
EPC	Energy Performance Certificate
Eps	Epsilon
ESS	European Statistical System
f_i	frequency of each group
$GBPN$	Global Buildings Performance Network
GI	Gini Index
GIS	Geographic Information System
$HVAC$	Heating, Ventilation, and Air Conditioning

i	data points
j	data points
LOF	Local Outlier Factor
lrd	local reachability density
$MinPts$	Minimum Points
N	number of cluster
n	data size
NN	Nearest Neighbors
s	standard deviation
$SEAI$	Sustainable Energy Authority of Ireland
SI	Silhouette Index
SSE	Sum of Squared Error
t	tonne (metric)
$TABULA$	Typology Approach for Building Stock Energy Assessment
$UBEM$	Urban Building Energy Modeling
x_i	group mid point

1. Introduction

The building sector accounts for a significant portion of the overall energy demand and greenhouse gas emissions of any country. Around 39% and 36% of the CO₂ emissions are associated with buildings in the United States and Europe respectively [1, 2]. According to the International Energy Outlook 2017 report, the commercial and residential building stocks in the United States account for around 35% of overall US energy consumption [2]. In Europe, buildings are responsible for 40% of overall energy consumption. As is evident from the statistics, buildings play a crucial role in the supply and demand of urban energy and therefore, possess considerable potential to achieve significant reductions in energy consumption and greenhouse gas emissions. These reductions are possible through the transformation of existing inefficient buildings into more efficient and sustainable forms which have the potential to improve resource usage and enhance system integration at an urban level.

One viable solution is to analyze the energy performance of buildings using the constraints of limited information available from the existing building stock; analysis can be accomplished using certain types of building energy modeling. The results of which can be

16 used to identify energy reduction measures that would in turn lead to a decrease in energy
17 consumption [3]. These energy models can help to analyze and compare different design
18 or retrofit design scenarios. In addition, these models can also aid in significantly improv-
19 ing the energy performance of buildings. Energy modeling at an urban scale often requires
20 building stock data, namely, geometry and building physical parameters. However, building
21 stock data collection is difficult at an urban scale due to lack of data and users' privacy
22 issues [3]. To bridge this gap, the building stock is commonly characterized using several
23 building-types that represent similar technical, operational and geometrical characteristics
24 of a large group of buildings [4].

25 A building stock is usually categorized as building typologies, reference buildings and
26 building archetypes [4]. The first category, *building typologies*, groups different buildings
27 using the criteria of building function, some of the examples include residential buildings,
28 universities, offices, and schools. The second category, *reference buildings*, is a concept for-
29 mulated by the European Union Energy Performance of Buildings Directive (EPBD) which
30 involves the development of *cost-optimal* frameworks, which can then be used to calculate the
31 cost-optimal levels of minimum energy performance requirements for buildings and building
32 elements [5]. All the member states in the EU are required to define reference buildings that
33 represent the average building stock in each member state (Commission Delegated Regu-
34 lation No. 244/2012) [6, 7]. The third category, *building archetypes*, uses geometric and
35 non-geometric parameters to classify the building stock. All buildings possessing similar pa-
36 rameters are grouped together and are termed as archetypes [8, 9]. The building archetypes
37 concept is most commonly used in energy modeling at the urban scale [10, 11]. So far, these
38 archetypes have been developed through the use of national survey data, which provides an
39 adequate overview of the national building stock. However, these archetypes become invalid
40 when identifying opportunities for deep retrofits at district/city scale [12, 13, 14, 15].

41 The building archetypes approach forms a part of bottom-up Urban Building Energy
42 Modeling (UBEM) that delivers a high level of accuracy with lower computational effort.
43 An archetype is usually defined as a virtual building that represents a number of buildings
44 sharing similar characteristics in the stock [16]. Because of the underlying difficulties in gath-
45 ering detailed information at an urban scale, the archetypes approach has become popular
46 in urban energy modeling as available information can be used to model similar buildings
47 [11]. The development of building archetypes is a complex task that involves gathering of
48 detailed building information at an urban scale.

49 Although existing databases and projects provide a valuable overview of a building stock,
50 the information is usually outdated and might lead to invalid energy consumption predic-
51 tions. Numerous studies and building stock databases exist in literature; these provide a
52 means to model the urban residential stock. For instance, databases such as ODYSSEE-
53 MURE [17], Eurostat [18], TABULA [19], CRB [20], ENTRANZE [21] and BPIE [22] devel-
54 oped building archetypes at a national level for different countries. The use of national scale
55 archetypes implies several assumptions for energy modeling at the urban or district scale.
56 For instance, studies such as [12, 13, 14, 15] model different urban areas in various parts of
57 the world using these national level archetypes but the analysis within these studies lacks
58 investigations at a fine temporal resolution. Other studies such as time of use surveys to

59 collect building stock data and predict energy consumption profiles [23, 24]. Although these
60 studies furnish the models with updated information, the underlying methodology tends to
61 be case specific, thus lacking the scalability often required in UBEM.

62 Various model assumptions born out of a lack of accurate data can cause significant
63 uncertainties in building energy models. Furthermore, no single archetype model can rea-
64 sonably represent the entire building sector of the same type in bottom-up models using
65 archetypes; more detailed classification of archetype models is often needed to avoid this
66 shortcoming.

67 Furthermore, these projects have significant associated uncertainties due to a lack of
68 knowledge and limited availability of statistical data. For instance, in TABULA, there
69 are assumptions made when defining the building classes, thermal properties, operational
70 characteristics and other properties. These projects cover national top-level archetypes only
71 and lack crucial information related to district scale or small-area building stocks.

72 Collection of input data forms one crucial step of the archetype development process but
73 the available building stock data is often incomplete and fails to match modeling standards.
74 Several inconsistencies often arise in such datasets as the building stock data is collected
75 through surveys; these include: human errors, incompleteness and coverage sampling. There-
76 fore, extracting useful, actionable, and interesting information from these data is extremely
77 difficult for the stakeholders. In this context, data-driven approaches that use data mining
78 and machine learning algorithms successfully provided solutions to extract useful informa-
79 tion from raw data [25, 26]. Over the past few years, data-driven approaches gained a lot
80 of attention in the field of building energy modeling [27]. However, previous research in
81 the area of building energy-related applications mostly focused on energy modeling, load
82 forecasting, energy prediction, and energy pattern profiling [28]. Other fields successfully
83 conducted numerous studies on data quality improvement and knowledge extraction but
84 limited work has been done in the building energy sector domain.

85 This paper uses data driven methods to improve the quality of data inputs required for
86 archetypes development. Some of the methods include data preprocessing, feature selection,
87 outlier detection, clustering methods. A data preprocessing method removes the inconsis-
88 tencies before the data is used with the data mining algorithms and data cleaning helps
89 improve algorithm performance. Similarly, the outlier detection process eliminates noise in
90 the data or removes the observations with exceptionally dis-similar information. A data
91 set for a given building stock contains many features or variables that do not influence the
92 archetypes development process. Optimal extraction of features would improve the quality
93 of energy modeling results. Feature extraction processes involve the extraction of the most
94 representative and useful variables from the data while clustering is a technique of assigning
95 a set of objects to the same group (called a cluster) so that the objects in a particular cluster
96 have similar values to each other and are significantly dissimilar to those in other clusters.
97 Therefore, the clustering process can help automated archetype generation rather than using
98 data mining algorithms.

99 There is a need to define these archetypes at various scales as national scale archetypes
100 often do not represent the different types of buildings existing at a local level. This pa-
101 per proposes a novel data-driven methodology for the development of multi-scale residential

102 building archetypes. The devised methodology integrates the building stock dataset avail-
103 able at different scales, for instance national, local authority, city and district. Furthermore,
104 the methodology uses data driven techniques to improve the quality of input data used
105 for archetype development. The previously developed archetype geometries coupled with
106 the parameters determined by using data-driven approaches to calculate annual energy use
107 (kWh) of buildings at a multi-scales. This process aids in providing a fine resolution to
108 the dataset and further nourishes it with the missing information. Furthermore, the im-
109 plemented methodology produces archetypes at the lowest possible scale, which enhances
110 the accuracy of urban energy simulations. As such, this helps eliminate the assumptions
111 associated with national level archetypes.

112 We apply the methodology to the building stock in Ireland and a comparative analysis
113 is performed at national, city, county and district scales. The proposed solution introduces
114 the following key features:

- 115 • Evaluation of existing approaches or projects of building archetype;
- 116 • Developing a generalized multi-scale building archetype methodology;
- 117 • Implementing the data-driven approaches for the characterization of building archetypes
118 and
- 119 • Comparison of the modeling results of multi-scale building archetype at BEM and
120 UBEM levels.

121 The development of multi-scale archetypes follows five steps. 1) initially involves data
122 collection from the existing residential building stock at multiple scales (National, local
123 authority, city and district). After data collection, 2) the segmentation approach further
124 divides the data to determine the number of archetypes represented by the building stock.
125 In step, 3) characterization, we describe the characteristics of archetypes using different
126 data-driven approaches. A quantification process detailed in step 4 determines the distri-
127 bution of archetypes using national building statistics data. Finally in 5), the last step, we
128 perform dynamic energy simulations on developed archetypes for the analysis and validation
129 of results.

130 This paper is structured as follows: Section 2 describes an overview of existing work
131 in archetypes development; Section 3 describes the proposed methodology, including an
132 explanation of the different steps of archetypes development; Section 4 states the results of
133 Irish case study followed by a discussion of the results. Section 5 concludes the research by
134 describing challenges and future work.

135 **2. Literature Review**

136 The building stock is broadly characterized into residential and non-residential build-
137 ings. Common examples of residential building stocks include houses and apartments. Non-
138 residential building stocks comprise industries, commercial complexes and others. Census

139 and survey data are the two most crucial resources for gathering information about a build-
140 ing stock [29]. Census data provides an essential source of statistical data ranging from
141 small areas (or neighbourhoods) to national and international scales. Survey data is the
142 statistical method through which additional sampling studies of individual units is carried
143 out in a population. In the United States, the Building Performance Database (BPD) is
144 the largest dataset that contains physical energy performance information of commercial
145 and residential buildings [30]. In Europe, the European Statistical System (ESS) developed
146 a new tool, "Census Hub" [31] that provides access to different national census databases.
147 Similarly, an extensive survey was conducted by the Buildings Performance Institute Eu-
148 rope (BPIE) that gathered information about the existing building stock across Europe to
149 identify potential ways to improve the energy performance of buildings. A Data Hub portal
150 was also released by BPIE to gather statistical data about the building stock characteristics
151 in EU [32]. Another project, ASsessment and Improvement of the EPBD Impact (ASIEPI)
152 for new buildings and building renovation, provided support in the identification of poten-
153 tial problems when implementing the EPBD in the EU [33]. Most of the member states
154 in Europe contain Energy Performance Certificate (EPC) databases that contain essential
155 information about their respective building stocks [34].

156 The building stock surveys provide additional information about buildings, including
157 usage patterns, technical characteristics and fuel usage. This kind of information helps
158 with energy analysis and modeling. Building stock surveys also help countries interested in
159 potential energy savings and CO₂ emissions. Other sources of building data include indi-
160 vidual billing data, sub-metering, energy certificates and geographical information systems
161 [35, 36, 3].

162 In 2013, the Global Buildings Performance Network (GBPN), published a report on
163 "*Comparing Data Quality & Collection*" [37] that described a comparative study of build-
164 ings' data robustness in four priority regions - Europe Union, China, India and the United
165 States. This report's result indicates that significant improvements are possible such as im-
166 proved data quality and availability after comparing European buildings data availability to
167 the data availability in the United States. The European Commission has made significant
168 investment in buildings' data collection and analysis projects such as ODYSSEE-MURE
169 [17], Eurostat [18], TABULA [19], ENTRANZE [21] and BPIE [22].

170 While the aforementioned databases and projects provide valuable data and an overview
171 of EU building stock, this data could not be used directly for archetype-based energy mod-
172 eling. The main issues include the lack of a physical description for the buildings and the
173 relatively outdated nature of the associated information. However, some databases such as
174 BPD are updated on a regular basis, and significant improvements have been made in the
175 past few years in terms of building data collection.

176 The work by Ballarini et al. implemented a simple statistical method for the interpre-
177 tation of data and removal of outliers [38]. However, the research presented in our paper
178 implements a data mining approach to improve the quality of input data for archetype de-
179 velopment. The data mining approach is a more appropriate approach for large datasets
180 when compared to simple statistical methods [28, 27]. Ballarini et al. further used a hierar-
181 chical clustering method for classification and thereby, the study also lacked a generalized

182 methodology for energy modeling of a large building stock [38].

Table 1: Literature review of existing residential building stock classification studies [16]

Paper	Concept			Scale				Segmentation			Context	
	Building typology	Reference Buildings	Building Archetypes	National	City	Regional	Districts	Dwelling Types	Year of Construction	Climate Zone		Others
[19]	X			X				X	X	X	X	EU
[39]		X				X		X		X	X	BR
[13]		X			X			X	X	X	X	PR
[14]			X		X				X		X	KW
[15]			X		X			X	X	X	X	LB
[29]			X	X				X	X	X	X	FGSU
[23]		X					X	X	X	X	X	CH
[12]	X			X					X		X	GR
[40]			X	X				X	X		X	IR
[9]			X	X				X	X		X	IR
[41]	X			X				X	X		X	DN
[12]	X			X				X	X	X	X	GR
[42]			X	X				X			X	DN
[43]	X			X					X	X	X	GR
[7]		X		X				X	X		X	EU
[44]	X			X				X	X		X	BSHC
[45]	X				X			X	X	X	X	IT
[46]		X		X				X		X	X	US

EU, Europe; BR, Brazil; PR, Portugal; KW, Kuwait; LB, Lisbon; FGSU, France, Germany, Spain and the United Kingdom; CH, China; GR, Greek; IR, Ireland; DN, Denmark; BSHC, Bulgaria, Serbia, Hungary, Czech Republic; IT, Italy; US, United State;

183 European and international level projects, such as, TABULA [47] and CRB [20] have
 184 developed archetypes that define the building stock [13]. Typology Approach for Building
 185 Stock Energy Assessment (TABULA) is one of the major initiatives at an integrated EU
 186 level to develop a database of different European building typologies. The database contains
 187 building typologies for 13 nations and consists of residential buildings classification, building
 188 types, energy consumption, possible energy savings and statistical data for buildings. An-
 189 other initiative by United States Department of Energy (DOE) involved the development of
 190 Commercial Reference Buildings (CRB) models. There are 16 building types represented by

191 the developed benchmarks that cover a total of 17 climate zones. Even though the projects
192 comprise a comprehensive database, the data is not updated on a regular basis. Hence,
193 the database lacks the information about any new potential archetypes. For instance, the
194 Irish building typologies in the TABULA database have remained the same since 2014. The
195 projects described above have significant associated uncertainties due to lack of knowledge
196 and limited statistical data availability. For instance, in TABULA, there are assumptions
197 made when defining the building classes, thermal properties, operational characteristics and
198 other properties [48].

199 The literature review summary for some of the studies on residential building stock
200 classification is shown in Table 1. Previous literature mostly covers the national top-level
201 archetypes and lacks crucial information related to district scale or small area building stocks
202 [49, 14, 13, 9, 16]. Furthermore, developed approaches are only tested on specific areas or
203 construction periods. The majority of the studies do not implement data-preprocessing
204 techniques, which are crucial for improving the quality of results. There is a need for a
205 generalized method of statistical analysis that can be used in any scenario. In addition,
206 there is a trade-off required between the time and data input needed for building stock
207 modeling. Alongside, urban energy modeling often requires identification of the level of
208 detail embedded in archetype formulation to achieve better results.

209 Most research efforts for the bottom-up engineering methods have focused on residential
210 buildings using archetypes. The majority of these models are tailored for residential buildings
211 in specific regions. For instance, studies by Famuyibo et al., Vasconcelos et al., Cerezo et
212 al. and Csoknyai et al. modeled the residential building stocks of Ireland, Portugal, Kuwait
213 and Eastern European countries respectively [9, 13, 14, 44]. A few of these studies integrate
214 the building stock models with a GIS platform for acquisition and expression of data. For
215 instance, a study by Li et al. developed urban residential reference buildings using clustering
216 analysis of satellite images [23].

217 Nearly all of these studies work with national level archetypes. As the energy policy
218 decisions require more detailed information to address future planning interventions, building
219 stock models need to be developed for a finer temporal and spatial resolution.

220 This paper addresses the aforementioned gaps identified through a generic data driven
221 approach. The devised methodology implements the granular level concept in the develop-
222 ment of archetypes. The methodology also implements detailed in-depth concepts from data
223 science to improve accuracy associated with energy modeling inputs.

224 **3. Methodology**

225 This paper assesses the impact of considering different scales of archetypes when char-
226 acterizing the residential building stock. To examine the different scale of archetypes, we
227 propose a generic approach to formulate multi-scale archetypes for residential buildings.
228 The approach uses data driven techniques to improve the quality of input data used for
229 archetype development. The archetype geometries developed previously [19], coupled with
230 the parameters determined by using the data-driven approach are used to calculate annual
231 energy use (kWh) of buildings at multiple scales.

232 The development of multi-scale archetypes follows the five steps outlined in Figure 1.
 233 The initial step involves data collection from the existing residential building stock at multi-
 234 scale (National, local authority, city and district). After data collection, the segmentation
 235 approach further divides the data to determine the number of archetypes represented by
 236 the building stock. In the next step, characterization, we describe the characteristics of
 237 archetypes using different data-driven approaches. A quantification process determines the
 238 distribution of archetypes using national building statistics data. Finally, in the last step, we
 239 perform dynamic energy simulations on developed archetypes for the analysis and validation
 240 of results. The following sections describe the individual steps of the methodology in further
 241 detail.

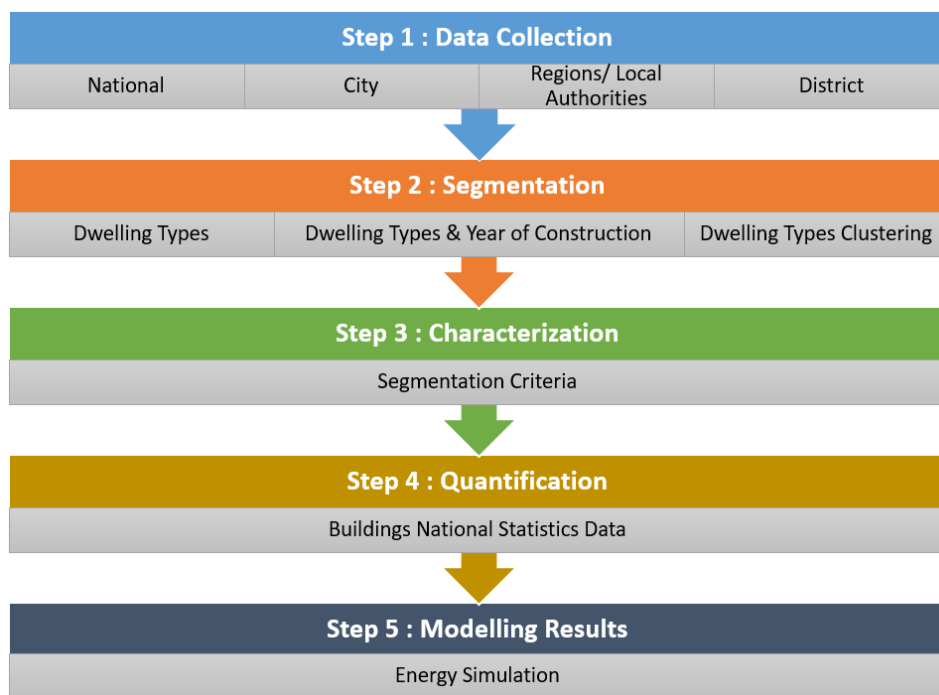


Figure 1: Overarching methodology for multi-scale building archetypes development

242 3.1. Data Collection

243 Data collection is the initial and most crucial step in the archetype development process.
 244 Building data is often extracted from the existing building stock for the development of
 245 building archetypes and requires geometric and non-geometric information related to any
 246 building. Geometric data includes building shape, dwelling type, building envelope, number
 247 of floors, walls, and windows, while non-geometric building data includes envelope U-values,
 248 construction assemblies and HVAC systems. In this paper, the residential building geometric
 249 and non-geometric data is collected at multiple scales from existing available building stocks
 250 or surveys such as:

- 251 • National scale : The building stock that represents the whole country.

- 252 • City scale: City building stock group of regions/ local authorities buildings.
- 253 • Regional/ local authorities scale: Geographical division of a city into different areas
254 where each local authority constitutes a number of districts.
- 255 • District scale: A district constitutes a group of small areas or neighbourhoods.

256 *3.2. Segmentation*

257 After the collection of data, the next step is the segmentation process that investigates
258 the number of archetype buildings required to represent the residential building stock at
259 multiple scales. There are various criteria for segmentation of the building stock such as
260 by dwelling type, construction year, climate zone [4]. Dwelling type and construction year
261 criteria are considered when performing segmentation in this research. As the main focus
262 of this research is to accurately calculate the energy demand of one region, the climate zone
263 criterion is not taken into account as the area in question (constituting several districts)
264 can only be associated with one climate zone. A data driven approach is used instead of a
265 simple statistical technique for the segmentation and characterization of the building stock.
266 Over the past few years, data driven approaches have emerged as the most common way
267 to perform analysis on large, unstructured and noisy data [28, 27]. This is mainly due to
268 the fact that data driven approaches allow to remove the anomalies in the data and hence,
269 produce models with high accuracy. Segmentation is performed using the criteria detailed
270 in the following sub-sections:

271 *3.2.1. Dwelling Types*

272 Building dwelling types are defined by size and layout of the building. The classification
273 of buildings plays a crucial role in energy modeling because energy use varies according
274 to the topology. For instance, detached, semi-detached or terraced houses have different
275 heating requirements and energy use.

276 *3.2.2. Year of Construction*

277 A building's year of construction, also known as the age of the building, impacts the
278 building's energy use. According to European Commission, currently, more than 35% of the
279 European Union's buildings are over 50 years old and almost 75% of the building stock is
280 energy inefficient. At the building stock level, this shows that year of construction is a key
281 parameter because of advancements in construction technology; older buildings normally
282 consume more energy when compared to newer buildings per unit area [1].

283 *3.2.3. Dwelling Types Clustering*

284 Clustering is an unsupervised machine learning technique for assigning a set of objects
285 to the same group (called a cluster), so that objects in a particular cluster have similar
286 values to each other when compared to those in other clusters. This research implements
287 unsupervised machine learning approaches to identify clusters of similar buildings based on
288 the characteristics of each dwelling type.

289 *3.3. Characterization*

290 This work examines the thermal properties of each building archetype including the
 291 construction material, usage patterns, and building systems [3] and subsequently extracts
 292 this information from a characterization process on the basis of selected segmentation criteria
 293 using a data-driven approach as opposed to a simple statistical approach (Figure 2). The
 294 building archetypes for UBEM require geometrical and non-geometrical data inputs for
 295 characterization of each building stock. These inputs are also necessary for an energy model
 296 that includes mainly construction type and glazing (U-Values), heating and cooling system,
 297 internal loads and occupancy profiles. The following sections describe the methodology for
 298 characterization of archetypes.

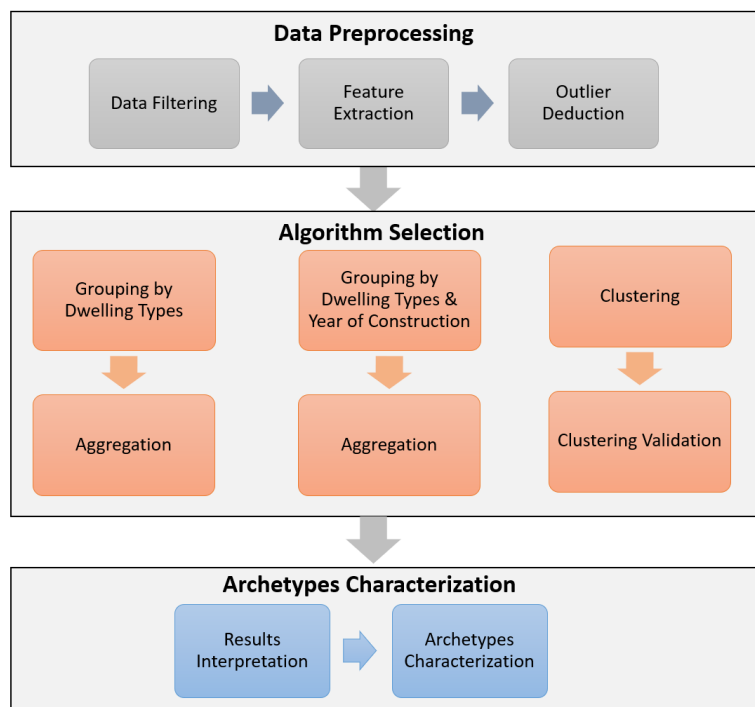


Figure 2: Characterization sub-methodology to determine the characteristics of building by using a data-driven approach

299 *3.3.1. Data Pre-processing*

300 This step involves implementing standardized procedures to remove inconsistencies in
 301 the data. Building stock data is usually collected from statistical surveys or questionnaires.
 302 Data collection through surveys is often prone to inconsistencies such as missing, duplicate,
 303 incomplete data. It is crucial for computational performance to remove these inconsistencies
 304 before the data is used as input to the data mining algorithms. Data filtering forms the
 305 first step in a data pre-processing procedure. In this paper, the implemented data filtering
 306 techniques include data cleaning, integration, transformation and discretization [50]. Data
 307 cleaning aids in resolving the inconsistencies in data such as missing values, outliers and

noisy data. Data integration techniques combine different data sources. Data transformation techniques change the nominal data to numeric data type for clustering algorithm implementation. Data discretization converts the construction years to age bands.

A feature extraction method follows the data filtering procedure. In the building stock data, there are many features or variables that do not influence the archetypes development process. Optimal extraction of features improves the quality of results when characterizing archetypes [9]. The feature extraction process involves extracting the most representative and useful variables from the data. As historical datasets often consist of irrelevant or redundant variables, this procedure improves the quality of data by selecting the most appropriate inputs. This process reduces the dimensionality of the model inputs resulting in significantly lower computational loads. Feature extraction is generally done using engineering or statistical methods [51, 52, 53]. Engineering methods make use of experts' interpretations and existing practices in the literature. On the other hand, statistical methods employ statistical or data mining methods such as regression analysis, neural networks. We implemented the feature selection using engineering methods in this study because a number of studies exist in literature; these have developed techniques to identify a minimum number of crucial features required to facilitate the energy modeling process [9, 54].

The last step in data pre-processing is the outlier detection procedure. The building stock data obtained through surveys often contain volumes of raw data that may include some outliers or incorrect information that should be treated before the data is used for archetype development.

It is important to remove noisy data before the implementation of clustering techniques on a large dataset so the outlier detection process eliminates the noise in the data or removes the observations with exceptionally dis-similar information. Outlier detection techniques are broadly categorized as: distance-based, density-based and Local Outlier Factor (LOF). In this paper, we used the LOF algorithm for detecting the outliers because LOF compares the outliers to their local neighborhoods rather than the global data distribution. Breunig et al. proposed the LOF outlier algorithm to identify the outlier data points utilizing the nearest neighbors (to calculate the degree to which an object is an outlier) [55]. LOF of an object i is the average ratio of local reachability density (lrd) of i and those of j Nearest Neighbors (NN). LOF can be computed using Equation (1), where, i and j are two data points.

$$LOF_{MinPts}(i) = \frac{\sum_{j \in N_{MinPts}(i)} \frac{lrd_{MinPts}(j)}{lrd_{MinPts}(i)}}{||N_{MinPts}(i)||} \quad (1)$$

3.3.2. Algorithm Selection

Algorithm selection is based on the choice of segmentation criteria. When segmenting by dwelling type, the first step groups data on the basis of dwelling types and subsequently performs the aggregation process. Similarly, the dwelling type and year of construction segmentation process involves grouping the data on the basis of dwelling types and year of construction and then performing the aggregation process that involves arithmetic or geometric mathematical operations. The formula for finding the mean (\bar{x}) and standard deviation (s) for grouped data are shown in Equation (2) and Equation (3) respectively.

$$\bar{x} = \frac{\sum x_i f_i}{n} \quad (2)$$

$$s = \frac{\sqrt{\sum x_i^2 f_i - \frac{\sum (x_i f_i)^2}{n}}}{n - 1} \quad (3)$$

347

348 where n is data size, x_i denotes the group mid point and f_i denotes the frequency of each
 349 group.

350 In the dwelling types clustering segmentation process, the k-means algorithm is used for
 351 clustering building stock data as this algorithm is considered to be the best approach for
 352 archetypes development when compared to others clustering algorithms [56]. K-means is
 353 the most common unsupervised partitional classification algorithm used to solve the clus-
 354 tering problem [57]. Each cluster is represented by the mean of the cluster with the aim
 355 of dividing the observations into k clusters where each observation belongs to a respective
 356 cluster (center point). The objective is to minimize the sum of the distances of the points
 357 to their respective centroid. The most common distance definition is the Sum of Squared
 358 Error (SSE) minimization function, also known as the Euclidean distance. The objective
 359 function is given by Equation (4)

$$C = \sum_{i=1}^K \sum_{x \in C_i} \|x - c_i\|^2 \quad \text{Euclidean distance} \quad (4)$$

360 where c_i is the mean of the n data points in cluster C_i . The scalability and simplicity of this
 361 approach is a key advantage of the K-means clustering technique when compared to other
 362 algorithms. K-means has serious limitations when the data contains outliers and clusters
 363 are of different sizes and densities.

364 Internal validity indices measure the validity of clustering results and calculate the prop-
 365 erties of resulting clusters, such as compactness, separation and roundness. The most com-
 366 mon validity indices are Silhouette Index (SI) [58], Davies Bouldin Index (DBI) [59] and
 367 Gini Index (GI) [60].

368 The silhouette index of a cluster value is a measure of the ratio between separation
 369 and compactness. The calculation is based on the silhouette width of their cluster objects.
 370 Mathematically, the silhouette index for each object i can be represented by Equation 5.

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (5)$$

371 where $a(i)$ is the average dissimilarity between i from all other objects within the same
 372 cluster, and $b(i)$ is the lowest of the average dissimilarity between i and objects among other
 373 clusters. The silhouette index (SI) is a normalized index range between -1 to 1 and a value
 374 close to 1 is always good for clustering.

375 The cluster distribution performance measure is evaluated using Gini Index (GI) which
 376 is a measure of statistical dispersion or cluster competence. A lower value of GI suggests an
 377 equal distribution while a higher value of GI indicates an unequal distribution,

378 The Davies Bouldin Index (DBI) is the ratio of *within cluster distances* to *between cluster*
 379 *separation*. Therefore, the ideal case for clustering is when DBI is smaller; meaning cluster
 380 are well compacted and separated. The DBI is as defined in Equation 6 where N is the
 381 number of clusters,:

$$C_{DBI} = \frac{1}{N} \sum_{i=1}^N D_i \quad (6)$$

382 where N is the number of clusters and D_i is defined as in Equation 7:

$$D_i = \max_{j \neq i} R_{ij} \quad (7)$$

383 R_{ij} , the similarity measure between clusters C_i and C_j , is defined as:

$$R_{ij} = \frac{C_i + C_j}{D_{ij}} \quad (8)$$

384 D_{ij} is the separation between the i^{th} and the j^{th} clusters.

385 3.3.3. Archetypes Characterization

386 The entire process concludes with the interpretation of aggregation and clustering re-
 387 sults. Each aggregated value and cluster centroid value represent the characteristics of one
 388 particular building archetype. All the variables selected in the feature extraction phase rep-
 389 resent the building's physical properties such as construction material, ventilation systems,
 390 glazing (U-Values), heating and cooling systems.

391 3.4. Quantification

392 The quantification step determines the number of buildings classified by each building
 393 archetype. To aggregate the result, a parameter called weighting coefficient is assigned to
 394 each archetype building. Generally, the number of buildings are quantified through the
 395 use of buildings' national statistics or census data [61]. For example, R Arababadi used
 396 BRE's domestic and non-domestic fact files data for UK energy studies [62]. E. Mata used
 397 the national stock of France, Germany, Spain and the UK for building-stock aggregation
 398 through archetype buildings [4]. Similarly, the Irish Census of 2016 provides details for a
 399 total of 2,003,645 buildings [63]. The entire building stock dataset comprises information
 400 about the types of dwellings, year of construction, building energy ratings, fuel used for
 401 heating or cooling and occupancy levels.

402 For instance, Ireland's Census 2016 data provides details of the Irish housing stock
 403 with statistics about buildings, dwelling types, construction years, domestic building energy
 404 ratings, fuel used for heating or cooling and occupancy level.

405 Section 3.5 is the last and the most crucial step in the overall process. At this point,
 406 the data is collected, segmented, characterized and quantified using the preceding steps and
 407 forms the inputs to energy simulation models in order to obtain the energy performance
 408 indicators.

409 3.5. Modeling Results

410 Modeling constitutes the final step of the devised methodology. The archetype build-
411 ings are modeled using energy simulation software, for example EnergyPlus or IES VE. The
412 results are then analyzed to identify the impact of bottom-up energy modeling at a large
413 scale. Energy modeling of buildings is usually classified into two categories based on differ-
414 ent modeling levels: Building Energy Performance Simulation (BEPS) and Urban Building
415 Energy Modeling (UBEM). The BEM approach is suitable for implementation at the indi-
416 vidual building scale while the UBEM approach covers the entire district or urban level. In
417 this paper, modeling is performed at both BEM and UBEM levels to compare the difference
418 in modeling results.

419 The developed archetypes are modelled in the BEPS engine in order to determine an
420 energy consumption total and demand profile. The BEPS engine requires building geometry,
421 i.e. the shape and proportions of the building, that is extracted from existing available
422 building stock, for instance, DOE and TABULA. The only available source of building
423 information at a large scale is through the existing national building stock data. For a
424 bottom-up model of the building stock, errors at the individual building level are higher
425 than at the aggregate level as inaccuracies within specific instances tend to average out at
426 the aggregate level. Similarly, building physics parameters for the BEPS engine are extracted
427 from the data-driven characterization process discussed in previous section 3.3.

428 In this paper, EnergyPlus is used as a BEPS engine for modeling dynamic energy de-
429 mand of archetypes. The approach uses information extracted from a building stock using
430 data driven methods. A study by Egan et al. established that it is possible to simulate the
431 residential buildings using EnergyPlus with a limited number of carefully defined paramete-
432 rs. Archetypes are mostly used for large scale energy modeling and EnergyPlus is quite
433 frequently used to model and simulate benchmark buildings for large scale energy modeling
434 or performing complex parametric simulations [64].

435 The developed archetypes and the associated modeling results are further analyzed using
436 GIS visualization maps for large scale building energy modeling. GIS information is useful
437 for understanding the energy demand of a large area.

438 4. Results and Discussion

439 The main objective of this paper is to develop and compare archetypes at multiple scales
440 and evaluate the difference in energy demand predictions in one region by using the multi-
441 scale archetypes approach. A case study of the Irish building stock is chosen to establish
442 the differences in the calculated energy demand at different scales.

Table 2: Distribution of dwelling types data in the Irish EPC dataset at multi-scale Ireland (National), Dublin (City), Dublin City (Local authority) and Dublin 1 (District)

Dwelling Types	Ireland	Dublin	Dublin City	Dublin 1
Apartment	4428	2252	914	261
Basement dwelling	193	109	70	28
Detached house	184034	15436	1128	13
End of terrace house	49755	19496	6313	227
Ground-floor apartment	39115	18899	6473	839
House	38868	3306	623	8
Maisonette	7138	3057	737	86
Mid-floor apartment	47379	30776	13098	3222
Mid-terrace house	97906	39614	15138	598
Semi-detached house	180756	48422	8122	59
Top-floor apartment	40405	18493	7156	1122
Total	689977	199860	59772	6463

443 Energy modeling at an urban scale is quite challenging as information about any in-
444 dividual building is often unavailable. This issue can be addressed by performing energy
445 modeling using archetypes to represent similar geometric and non-geometric parameters of
446 a large group of buildings. To date, archetypes for energy modeling have usually been de-
447 veloped and implemented at a national scale shown in Table 1. This study aims to develop
448 and compare the performance of multi-scale archetypes (national, city, local authority and
449 district). The results could establish the importance of local level archetypes which could
450 in turn be beneficial to urban planners and policy makers to accurately identify the en-
451 ergy demand of a region. Furthermore, the use of multi-scale archetypes could aid in the
452 identification of deep retrofit measures within any specific area of interest.

453 4.1. Data Collection

454 The archetypes development methodology at multi-scale was applied to the publicly
455 available Irish Building Energy Performance Certificate (EPC) data published by Sustain-
456 able Energy Authority of Ireland (SEAI). An EPC rating is the overall energy consumption
457 of a building on a graded scale. An EPC for individual houses contains a building’s energy
458 performance rating in terms of normalized primary energy consumption ($\text{kWh}/(\text{m}^2 \cdot \text{year})$)
459 with the rating varying on a scale from A1 to G1 (Figure 3). An A1-rated building has
460 the lowest energy consumption and CO_2 emissions. On the other hand, a G-rated building
461 is the highest energy consumption and CO_2 emissions. The Dwelling Energy Assessment
462 Procedure (DEAP) software is Ireland’s official method for calculating the EPC rating of
463 new and existing dwellings. The calculation requires a range of building related information
464 in addition to geometric properties that includes space heating type, water heating type,
465 ventilation type, and lighting.

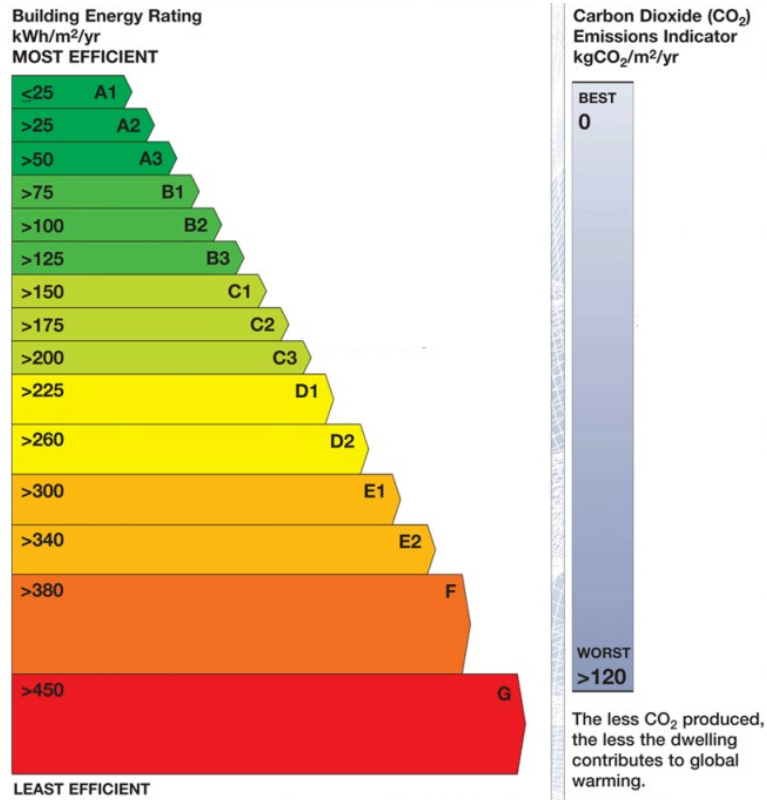


Figure 3: Irish EPC building energy and CO₂ emissions rating chart used to determine building energy performance

466 The Irish residential EPC data set contains more than 600,000 Irish buildings', each with
 467 203 variables including building physics, energy consumption, and associated CO₂ emissions.
 468 The data segregates the residential stock into various dwelling types, and does so at multi-
 469 scales such as Ireland (national), Dublin (city), Dublin City (local authority) and Dublin 1
 470 (district) shown in Table 2. Dublin County, Dublin City and Dublin 1 represent 29%, 9%
 471 and 1% of the EPC building stock data respectively. Currently, Irish EPC data covers 35%
 472 of entire residential building stock. Corrado et al. demonstrated that the EPC data used
 473 for building energy modeling is only slightly affected by data uncertainties [65].

474 4.2. Segmentation

475 Data segmentation follows the data collection process. For this particular case study, we
 476 performed the segmentation process initially with the dwelling types of which there are 11
 477 types of dwellings. To increase the level of detail and complexity, we combined the dwelling
 478 types with their respective construction age bands after identifying 9 such bands from the
 479 CSO dataset [63]. These age bands are based on the construction type and Irish building
 480 regulations [63]. We implemented clustering techniques to perform data segmentation and
 481 used individual building features, for instance, U-values, to identify the clusters within a
 482 particular dwelling type.

483 *4.3. Characterization*

484 The first step in the characterization process was data pre-processing, which included the
 485 enrichment and restructuring of data. We enriched and structured the EPC data through
 486 various methods such as replacing the missing values with average, removing useless variables
 487 or less frequent values using a standard deviation threshold. Aggregation and clustering was
 488 mostly applied to numerical values, so all the nominal values were converted to numerical.

489 While the dataset contains 203 variables, only variables that directly affected the energy
 490 dynamics were considered for the archetype development. We selected crucial variables
 491 using a feature extraction process. In this paper, an engineering method was considered
 492 for feature extraction, which employed existing literature to list out all the crucial features.
 493 For example, a study by Famuyibo et al. identified 24 variables that directly influence a
 494 building’s energy performance [9]. Similarly, Corrado et al. results demonstrated that that
 495 less than 10 variables, out of 129 variables, have a significant influence on the energy rating
 496 uncertainties [65]. Some of these features include building U-values, building areas, primary
 497 fuel source. The statistical summary of U-values (W/m^2K) is shown in Figure 4 .

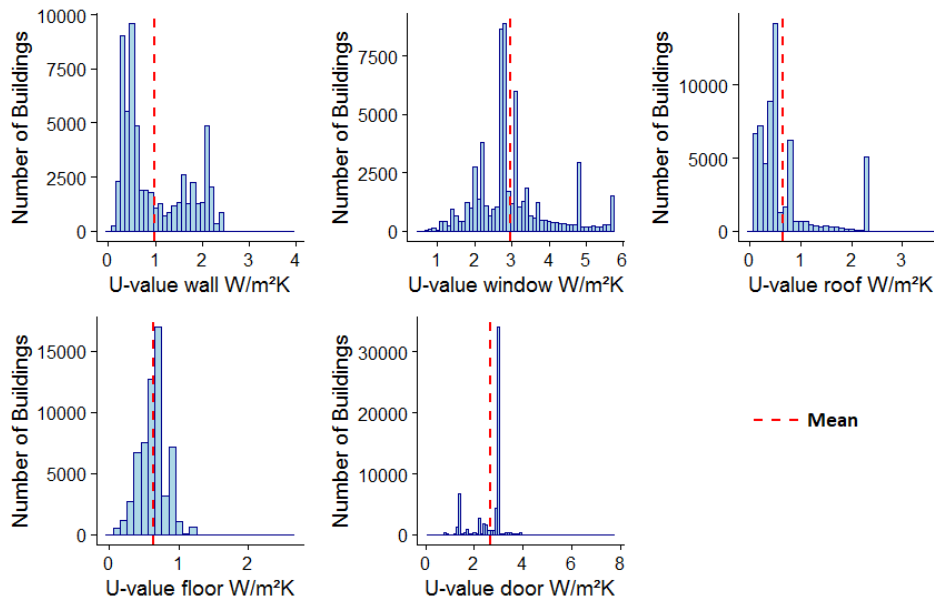


Figure 4: Histograms of building element U-values (W/m^2K) with the mean value of Dublin City EPC data set

498 We use the LOF algorithm to remove the outliers from the EPC dataset. For instance,
 499 the outliers in building window, wall roof, floor and door areas in EPC dataset are shown
 500 in Figure 5. The LOF algorithm is based on the distance function to measure the density
 501 of objects amongst each other. The Euclidean distance measure was used with the LOF
 502 algorithm for the case study.

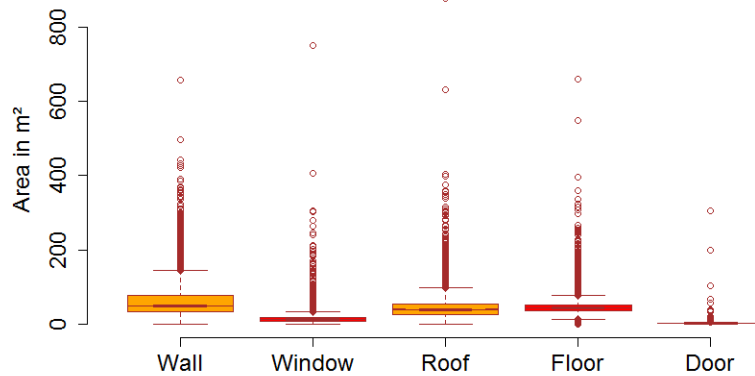


Figure 5: Boxplots of building elements area (m^2) shows outlier present in Dublin City EPC data set

503 Each individual building in the structured dataset contains its own set of values for
 504 different variables (features). The values for each variable were aggregated for buildings
 505 that belong to one particular segment (archetype). This aggregation resulted in a single set
 506 of values for the associated variables. For instance, there was a single set of values for each
 507 variable associated with one particular dwelling type. Aggregation of values for the segments
 508 created using dwelling types only and dwelling types with age bands was achieved using the
 509 arithmetic mean. These mean values for each dwelling type represent the characteristics of a
 510 unique building archetype. Therefore, 11 multi-scale building archetypes were identified on
 511 the basis of dwelling type segmentation and 99 multi-scale building archetypes are identified
 512 based on the dwelling type with age bands segmentation.

513 Aggregation for segments created through clustering of dwelling types was achieved using
 514 the k-means algorithm. Clusters are developed using different values of k (from 2 to 10) and
 515 the best k number of clusters are chosen for the 11 dwelling types with items distributed
 516 based on different validity indexes, namely, SI, GI and DBI as shown in Table 3. The process
 517 identified the maximum and minimum number of clusters to be 10 and 2 respectively while a
 518 total of 49 clusters (sum of individual clusters for each dwelling type) represent the building
 519 stock of Dublin city (Table 3). The k value is the number of clusters or archetypes present
 520 in each dwelling type. For example, clusters of the semi-detached houses have 4 ($k = 4$)
 521 classes that relate to 4 different building archetypes and these four archetypes represent
 522 48,474 buildings in the Dublin city local authority. The distribution of each cluster contains
 523 19704, 1492, 19257 and 8021 items respectively. This observation relates to the fact that
 524 there are four types of archetypes for semi-detached house and the results show that all the
 525 developed clusters are compacted and separated from each other. An identical approach
 526 was applied on all other dwelling types.

Table 3: K-mean number of clusters analysis of Dublin City for archetypes development

Dwelling Types	k	DBI	GI	SI	Clusters items distribution	Total
Mid-floor apartment	2	0.75	1.00	0.61	2206, 28540	30746
Top-floor apartment	2	0.97	1.00	0.42	6217, 12264	18481
Mid-terrace house	2	0.70	1.00	0.60	32313, 7572	39885
Semi-detached house	4	0.82	0.98	0.39	19704, 1492, 19257, 8021	48474
Detached house	3	0.25	0.88	0.68	7803, 2549, 827, 4230	15409
Maisonette	10	0.82	0.97	0.42	570, 281, 214, 199, 70, 860, 61, 261, 398, 113	3027
Ground-floor apartment	6	0.98	1.00	0.31	5245, 3104, 4642, 4737, 951, 248	18927
House	4	0.51	0.89	0.34	320, 1927, 40, 977	3264
Apartment	2	0.46	1.00	0.78	20, 2198	2218
End of terrace house	7	0.78	0.99	0.40	1611, 1918, 5078, 1228, 8675, 816, 185	19511
Basement Dwelling	7	0.69	0.90	0.39	8, 31, 8, 4, 27, 7, 22	107

527 4.4. Quantification

528 Quantification follows the data characterization process. The Irish Central Statistics
529 Office's (CSO) 2016 data was used for the quantification of archetypes; This research im-
530 plemented the quantification process of dwelling type segments for Dublin city at the local
531 authority scale. The CSO dataset provides the statistics of only 4 dwelling types, namely,
532 detached houses, semi-detached houses, terraced houses and flats or apartments. However,
533 11 dwelling archetypes were identified using the EPC dataset. Therefore, the 11 dwelling
534 archetypes identified were mapped into the available 4 dwelling types for the quantification
535 process (Table 4).

Table 4: Dublin City grouping EPC to Census dwelling types for quantification process

EPC Dwelling Types	Census Dwelling Types	Number of buildings	Percentage
Mid-floor apartment			
Top-floor apartment			
Apartment	Apartment	74537	35%
Maisonette			
Ground-floor apartment			
Basement Dwelling			
Semi-detached house	Semi-detached house	48881	23%
House			
Detached house	Detached house	10382	5%
Mid-terrace house	Terraced house	74446	35%
End of terrace house			
-	Not stated	3345	2%

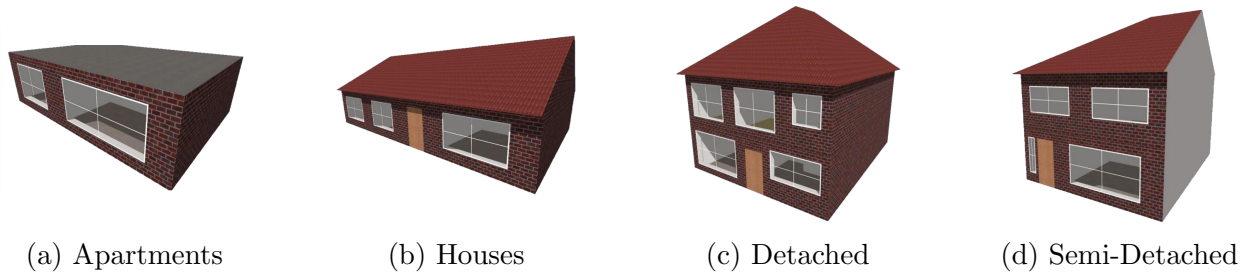


Figure 6: 3D model visualisations of Irish building stock for energy EnergyPlus simulation [66, 54]

536 *4.5. Modeling Results*

537 All the archetypes were developed at different scales using different segmentation meth-
 538 ods and were subsequently simulated in EnergyPlus. The building parameters identified
 539 in the characterization process were used as inputs to the energy simulation models. This
 540 process used previously developed archetype geometries coupled with the parameters deter-
 541 mined by the characterization process to calculate annual energy use (kWh) of buildings at
 542 multiple-scales.

543 EnergyPlus requires inputs including climate data, geometric, and non-geometric data.
 544 For this test case, we chose Dublin City EnergyPlus Weather (EPW) file as the climate
 545 input. We extracted the building’s geometric input using the existing literature on Irish
 546 buildings, for instance, Nue et al. developed comprehensive EnergyPlus geometric models
 547 to represent the Irish building stock shown in Figure 6 [66, 54]. The characterization process
 548 determined all non-geometric inputs including building physical parameters. The models are
 549 realized using the DesignBuilder interface to EnergyPlus and represent the dwelling types
 550 as per the census data of the Irish building stock. These models have already been tested
 551 and validated in term of kWh/(m²*year) with the Dwelling Energy Assessment Procedure
 552 (DEAP), Ireland’s official method for calculating the Building Energy Rating of new and
 553 existing dwellings [66]. These geometric models were used as base-case archetype model in-
 554 puts in the presented case study. Energy Modeling was performed at the individual building
 555 (BEPS) and large scale (UBEM) levels for detailed analysis.

556 *4.6. Building Energy Performance Simulation*

557 This section enlists the BEPS results for the developed multi-scale archetypes. As BEPS
 558 is mostly used for benchmarking or creating reference buildings using archetypes, therefore,
 559 the main goal is to evaluate the impact of using multi-scale archetypes on individual BEPS
 560 results.

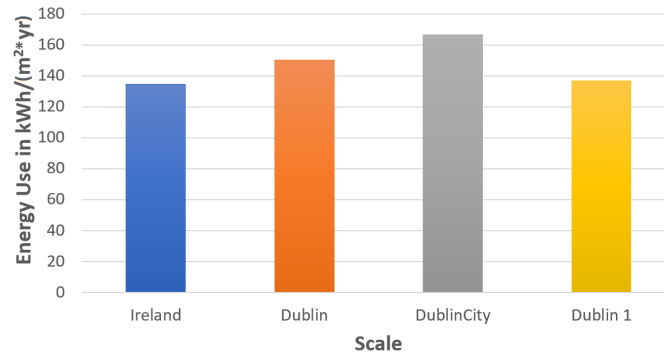


Figure 7: Annual site energy per total building area (kWh/(m²*year)) for a semi-detached house archetype at multiple scales using dwelling type segmentation

561 The annual site energy per total building area (kWh/(m²*year)) for semi-detached houses
 562 using dwelling type segmentation is shown in Figure 7. Significant differences in calculated
 563 energy were identified for the same dwelling when the calculations were performed using
 564 archetypes values from other scales. The differences in calculated energy use were found to
 565 be 16, 32 and 2 (kWh/(m²*year)) at the city, local authority and district scales respectively
 566 when compared to energy use at the national level. On a monthly basis, the differences
 567 in energy consumption becomes significant, especially during the winter season (Figure 8).
 568 End uses such as electricity, lighting, domestic hot water tended to stay the same at all
 569 archetype scales (Figure 9). Major differences occur in space heating demand as U-values
 570 (W/m²K) vary across archetypes at different scales.

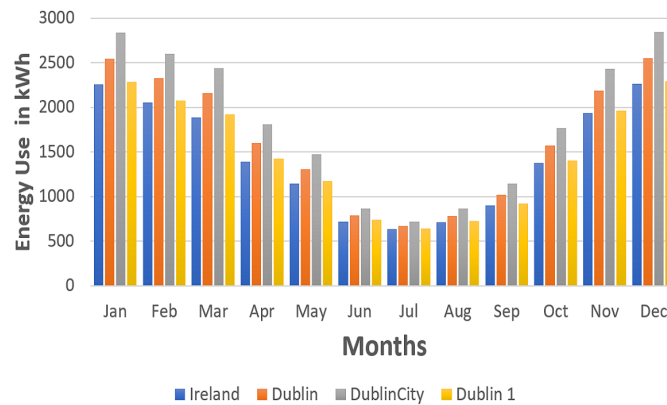


Figure 8: Monthly energy use in kWh of semi detached house archetype at multi-scale using dwelling type segmentation

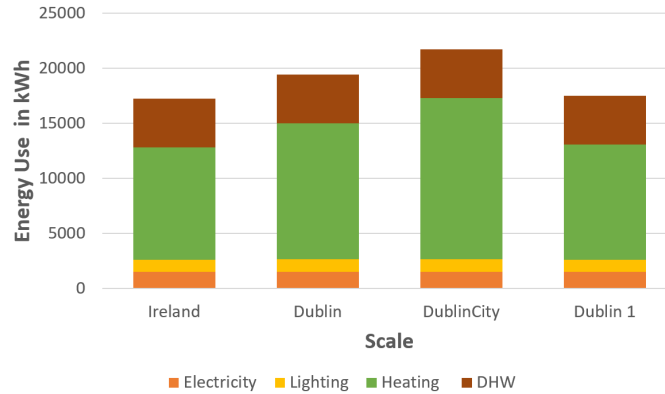


Figure 9: End use use in kWh of semi detached house archetype at multi-scale using dwelling type segmentation

571 An overall comparison of the different dwelling types show that the differences in maximum energy use were also significant for detached houses, houses and top-floor apartment shown in Table 5. Absolute differences in energy use were found to be 29, 57 and 5 (kWh/(m²*year)) for detached houses, semi-detached houses and top-floor apartment respectively. The Irish EPC rating scale has been formulated such that each grade has a scale size of 25 (kWh/(m²*year)). The use of national scale archetypes for BEPS could lead to an incorrect building energy performance calculation. For instance, when using national scale archetypes, the calculated energy consumption of a house dwelling type was 142 (kWh/(m²*year)) (equivalent to C1 rating based on SEAI Irish EPC Rating); this number increased by approximately 30% to 199 (kWh/(m²*year)) and the EPC rating shifted from C1 to C3 when detailed local level archetypes were used. Furthermore, the difference in the calculated energy consumption was found to be insignificant for apartments, basement dwellings and maisonettes because of minor differences in building physical parameters of these dwelling types at multi-scale. For instance, apartment window U-values are identified as 2.4, 2.3, 2.2 and 2.3 (W/m²K) for archetypes at Ireland, Dublin, Dublin city and Dublin 1 levels respectively (Table 5). These dwelling types cover less than 35% of entire Dublin City building stock (Table 4). However, more than 65% of the Dublin City building stock exhibit deviations when the parameters for different geographical scales are used.

Table 5: The comparison of dwelling types site energy per total building area in kWh/(m²*year) at multi-scale

Dwelling Types	Ireland	Dublin	Dublin City	Dublin 1
Apartment	109	108 ↓ (-1)	108 ↓ (-1)	107 ↓ (-2)
Basement Dwelling	130	135 ↑ (5)	133 ↑ (3)	133 ↑ (3)
Detached house	143	152 ↑ (9)	159 ↑ (16)	172 ↑ (29)
End of terrace house	138	144 ↑ (6)	161 ↑ (23)	138
Ground-floor apartment	114	115 ↑ (1)	118 ↑ (4)	117 ↑ (3)
House	142	157 ↑ (15)	175 ↑ (33)	199 ↑ (57)
Maisonette	113	113	119 ↑ (6)	114 ↑ (1)
Mid-floor apartment	113	112 ↓ (-1)	115 ↑ (2)	115 ↑ (2)
Mid-terrace house	145	152 ↑ (7)	168 ↑ (23)	147 ↑ (2)
Semi-detached house	135	150 ↑ (15)	167 ↑ (32)	137 ↑ (2)
Top-floor apartment	115	115	119 ↑ (4)	120 ↑ (5)

↑, site energy is more than National (Ireland) scale archetype; ↓, site energy is less than National (Ireland) scale archetype; () shows the difference with National (Ireland) scale archetype

589 Energy simulations were performed using archetypes created using the year of construc-
590 tion segmentation approach. The energy consumption was calculated at different archetype
591 scales for various construction age bands for the semi-attached house dwelling type. The
592 results establish the relationship between the age of a building and the energy consumed
593 (Figure 10). Furthermore, the calculated energy consumption differed significantly for differ-
594 ent archetype scales across all age bands (the differences being the highest between 1946 and
595 1980). It is crucial to note that year of construction significantly influence building energy
596 consumption. Therefore, archetypes implemented for urban energy modeling without the
597 consideration of construction year can lead to inappropriate modeling results.

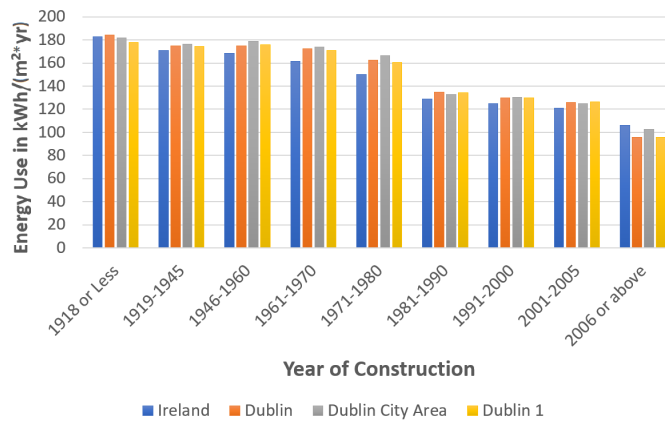


Figure 10: Comparison of semi detached house archetype energy use at different multi-scale by year of construction segmentation

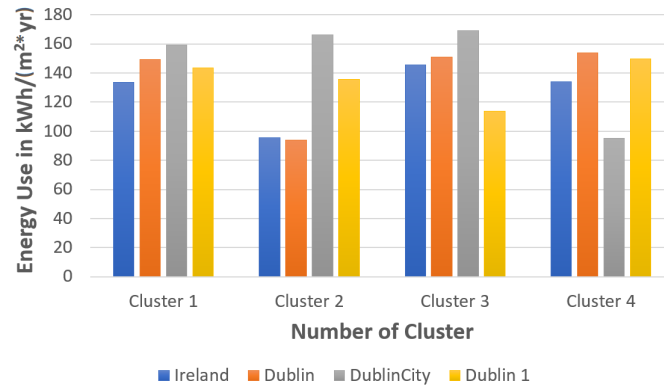


Figure 11: Comparison of semi detached house archetype energy use at different multi-scale by clustering segmentation

598 Archetypes developed through dwelling types clustering segmentation produced similar
 599 results as archetypes developed using the aggregation method. For instance, the differ-
 600 ences in the energy use (observed as a result of clustering segmentation) for semi detached
 601 houses range from 2 to 70 ($\text{kWh}/(\text{m}^2 \cdot \text{year})$) at the city, local authority and district scales
 602 respectively when compared to energy use at the national level, as shown in Figure 11.

603 4.7. Urban Building Energy Modeling

604 In this section, the developed archetypes were analyzed and used for urban building en-
 605 ergy modeling and quantification of data, therefore, this case-study dealt with only dwelling
 606 type and year of construction segmentation. The quantification of data helps in calculating
 607 total energy demand in the area by multiplying the archetype energy usage by the number
 608 of building types present in that area. The dwelling type clustering produces better results
 609 when limited quantification data is available at the urban scale.

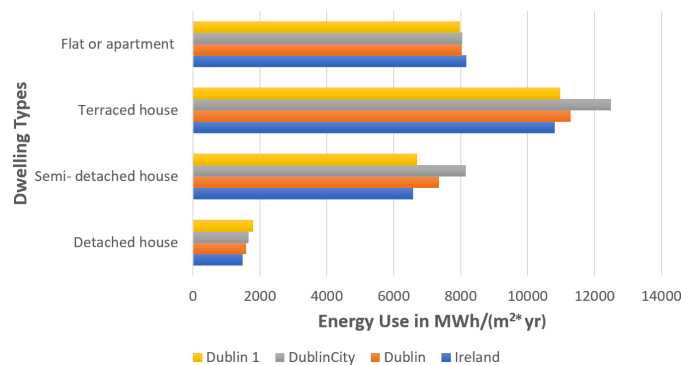


Figure 12: Different dwelling types energy usage at multi-scale of Dublin city

610 The building stock of Dublin city local authority was modelled at the regional/local
 611 authority scale using dwelling type segmentation. The calculated energy consumption values

612 were found to be 27011, 28224, 30324 and 27405 in MWh/(m^2 *year) at Ireland (National),
613 Dublin County (regional), Dublin City (city) and Dublin 1 (district) scales. The results
614 indicated a difference of 3313 MWh/(m^2 *year) in the calculated energy consumption when
615 national scale (Ireland) archetypes were used at local authority (Dublin City) scale. This
616 difference clearly indicates that the urban planners or policy makers neglected an equivalent
617 of ≈ 1599 tonnes of t CO₂/kWh produced from this region in their analysis. The results
618 also signify that the selection of an archetype for a particular scale introduces significant
619 differences in the predicted values of energy consumption and emissions. At an urban scale,
620 these differences often escalate, which leads to inappropriate results. The comparison of
621 predicted energy consumption for different dwelling types at the Dublin city scale is shown
622 in Figure 12. The consumption values were predicted using archetypes developed at the
623 four scales. The differences were highly significant for terraced and semi-detached houses
624 (Figure 12).

625 The results presented in the previous sections indicate that year of construction plays an
626 important role in building efficiency. Therefore, the building stock of the Dublin city local
627 authority was also modelled using the year of construction segmentation.

628 For better visualization of results, we created a GIS map of Dublin city and embedded
629 energy consumption values for the entire building stock in the map. The map was generated
630 for semi-attached houses that use semi-detached archetype using year of construction seg-
631 mentation. During the data mapping process, census data was used for small areas energy
632 demand calculation.

633 The detailed mapping of GIS-Based residential building energy modeling at Dublin City
634 local authority scale has been recently published in [67]. The map indicated almost similar
635 results to those devised from BEPS section 4.6. As evident from Figure 13, the energy map
636 created using the national scale archetypes was significantly different from the energy maps
637 created using the Dublin (City), Dublin City (Local Authority) and Dublin 1 (District) scale
638 archetypes. Therefore, archetypes developed at the national scale do not represent small
639 areas effectively and are not suitable when performing energy simulations at the district scale.
640 The results also indicate that maps using Dublin, Dublin City and Dublin 1 scale archetypes
641 are quite similar. This establishes the fact that accurate modeling of an urban area requires
642 archetypes at city scale or lower resolution. Urban planners often use GIS mapping to
643 identify areas that require energy retrofits and therefore, selection of an archetype scale is
644 crucial for accurately representing energy consumption variations at an urban level.

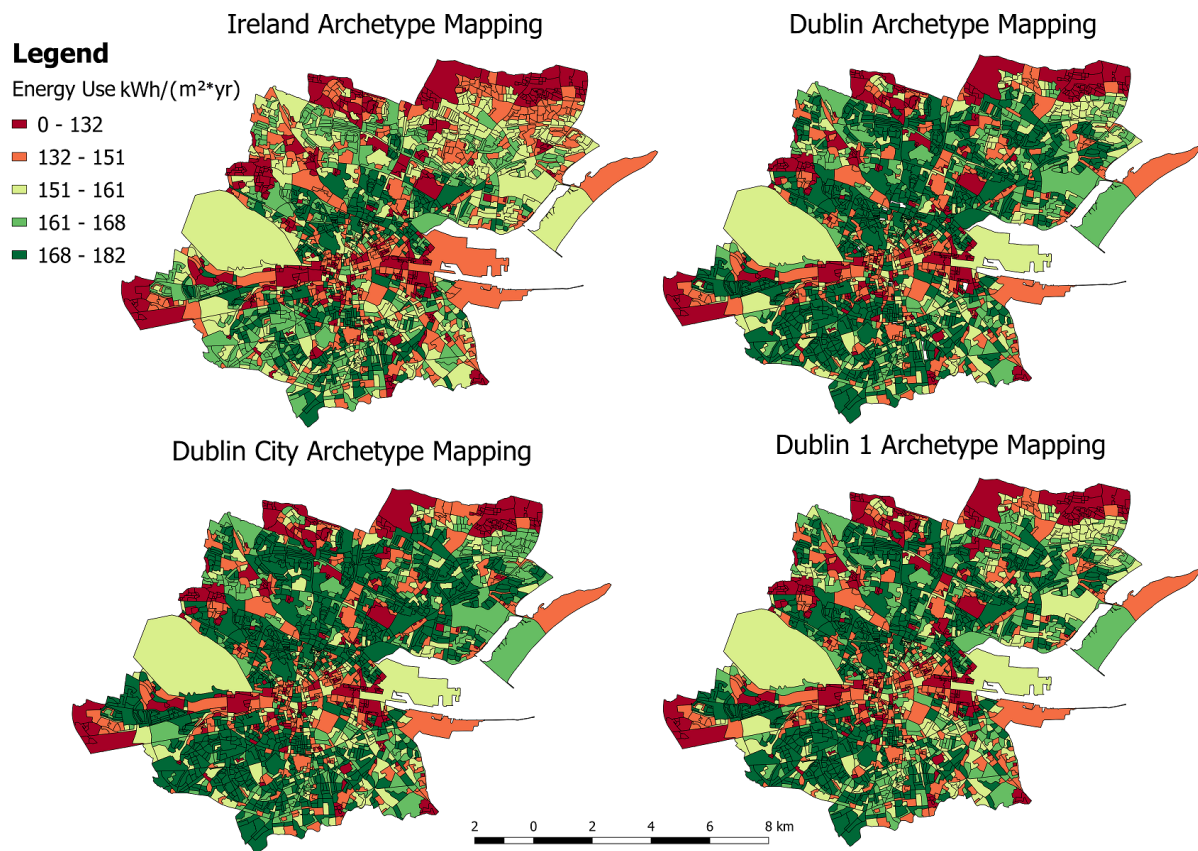


Figure 13: Comparison of Dublin city Local authority average annual small area energy usage in (kWh/(m²*year)) by mapping semi-detached archetype using year of construction segmentation at multi-scale

645 5. Conclusions and Future Work

646 At an urban scale, the implementation of energy efficiency measures requires appropriate
 647 modeling of the building stock. Archetype development can help to reduce the modeling
 648 effort and computational time. The research conducted in this paper proposes a general-
 649 ized methodology to develop archetypes of different dwelling types which will allow for an
 650 extended and a detailed analysis of the energy performance of buildings. The proposed
 651 archetype development methodology uses different data-driven approaches. The archetypes
 652 segmentation process is based on the dwelling type, year of construction and clustering
 653 method. The results show that bottom-up energy modeling at large scale could be further
 654 improved by using multi-scale building archetypes. This will assist the local authorities, city
 655 planners and energy policymakers to analyze granular level building energy performance,
 656 that further helps to improve sustainable energy policy decisions.

657 We also evaluated the methodology using a comparative study of developed archetypes
 658 at national, city, county and district scales. The approach leveraged Irish buildings and

659 the results indicate that archetypes developed at the national scale produce significantly
660 different modeling results when used at a local scale.

661 A major problem in UBEM simulations is uncertainty associated with the characteristics
662 of building archetypes. A lack of thermal properties and limited access to building energy
663 audit data can significantly increase uncertainty levels. These uncertainties could lead
664 to errors in UBEM simulations. The devised methodology will help in improving energy
665 modeling results and reduce uncertainty at an urban scale by using the developed data-
666 driven approach.

667 The results further indicate that the selection of an archetype scale introduces significant
668 differences in the calculated values of energy consumption and emissions. These multi-scale
669 archetypes could be further used as benchmarks or reference buildings to evaluate various
670 energy savings and energy efficiency strategies at a local area level. This work will also aid in
671 developing energy policy and will inform retrofit standards for future residential buildings.
672 The identified building archetypes can also be used as a guideline for construction of new
673 buildings or standard assessment methodologies to improve building performance at a large
674 scale. Furthermore, the developed archetypes would aid in the implementation of retrofit
675 strategies on existing buildings at an urban scale.

676 The results achieved by using the proposed methodology could be further improved
677 by using more detailed building physics and quantification data. Further research work
678 is required to integrate the dynamic occupancy behavior to enhance the modeling results.
679 As the methodology presented here is only tested on residential buildings, further research
680 could be extended to include commercial buildings. The methodology presents an auto-
681 matic process to generate up-to-date archetypes by linking the latest building stock. Future
682 work will develop a tool for the proposed methodology that can integrate with the latest
683 building stock to develop archetypes at multi-scale and thereby, automate the process of
684 archetypes simulation based on inputs gathered in the characterization process. The bot-
685 tom up engineering building stock models do not consider uncertainties associated with the
686 large building stock. Stochastic building energy stock models can handle these uncertainties
687 and would be a valuable addition to the presented research.

688 **6. Acknowledgements**

689 This publication has emanated from research supported in part by a research grant from
690 Science Foundation Ireland (SFI) under the SFI Strategic Partnership Programme Grant
691 number SFI/15/SPP/E3125. The opinions, findings and conclusions or recommendations
692 expressed in this material are those of the author(s) and do not necessarily reflect the views
693 of the Science Foundation Ireland.

694 **References**

- 695 [1] Energy for Europe by European Commission, <https://ec.europa.eu/energy/en>, [Online; accessed
696 10-July-2019].
697 [2] Building and Climate Change by Environmental and Energy Study Institute, <http://www.eesi.org/>,
698 [Online; accessed 10-July-2019].

- 699 [3] C. F. Reinhart, C. C. Davila, Urban building energy modeling—a review of a nascent field, *Building and*
700 *Environment* 97 (2016) 196–202.
- 701 [4] É. Mata, A. S. Kalagasidis, F. Johnsson, Building-stock aggregation through archetype buildings:
702 France, germany, spain and the uk, *Building and Environment* 81 (2014) 270–282.
- 703 [5] EU, Directive (EU) 2018/844 of the European Parliament and of the Council of 30 may 2018 amending
704 Directive 2010/31/eu on the energy performance of buildings and Directive 2012/27/eu on energy
705 efficiency, *Official Journal of the European Union* 61.
- 706 [6] EU, Commission Delegated Regulation (EU) no 244/2012 of 16 January 2012 supplementing Directive
707 2010/31, EU of the European Parliament and of the Council on the energy performance of buildings.
- 708 [7] I. Ballarini, S. P. Corgnati, V. Corrado, Use of reference buildings to assess the energy saving potentials
709 of the residential building stock: The experience of tabula project, *Energy policy* 68 (2014) 273–284.
- 710 [8] A. Galante, M. Torri, et al., A methodology for the energy performance classification of residential
711 building stock on an urban scale, *Energy and Buildings* 48 (2012) 211–219.
- 712 [9] A. A. Famuyibo, A. Duffy, P. Strachan, Developing archetypes for domestic dwellingsan irish case study,
713 *Energy and Buildings* 50 (2012) 150–157.
- 714 [10] C. C. Davila, C. F. Reinhart, J. L. Bemis, Modeling boston: A workflow for the efficient generation
715 and maintenance of urban building energy models from existing geospatial datasets, *Energy* 117 (2016)
716 237–250.
- 717 [11] J. Sokol, C. C. Davila, C. F. Reinhart, Validation of a bayesian-based method for defining residential
718 archetypes in urban building energy models, *Energy and Buildings* 134 (2017) 11–24.
- 719 [12] I. Theodoridou, A. M. Papadopoulos, M. Hegger, A typological classification of the greek residential
720 building stock, *Energy and Buildings* 43 (10) (2011) 2779–2787.
- 721 [13] A. B. de Vasconcelos, M. D. Pinheiro, A. Manso, A. Cabaço, A portuguese approach to define reference
722 buildings for cost-optimal methodologies, *Applied Energy* 140 (2015) 316–328.
- 723 [14] C. Cerezo, J. Sokol, S. AlKhaled, C. Reinhart, A. Al-Mumin, A. Hajiah, Comparison of four building
724 archetype characterization methods in urban building energy modeling (ubem): A residential case study
725 in kuwait city, *Energy and Buildings* 154 (2017) 321–334.
- 726 [15] C. S. Monteiro, A. Pina, C. Cerezo, C. Reinhart, P. Ferrão, The use of multi-detail building archetypes
727 in urban energy modelling, *Energy Procedia* 111 (2017) 817–825.
- 728 [16] C. SOUSA MONTEIRO, C. CEREZO, A. PINA, P. FERRAO, A method for the generation of multi-
729 detail building archetype definitions: Application to the city of lisbon, *Proceedings of International*
730 *Conference CISBAT 2015 Future Buildings and Districts Sustainability from Nano to Urban Scale*
731 (2015) 901–906doi:10.5075/epfl-cisbat2015-901-906.
732 URL <http://infoscience.epfl.ch/record/213443>
- 733 [17] ODYSSEE and MURE data bases, European-Union, <http://www.odyssee-mure.eu>, [Online; accessed
734 10-July-2019].
- 735 [18] Eurostat by Statistical office of the European Union, european-union, <http://ec.europa.eu/eurostat>, [Online; accessed 10-July-2019].
- 736 [19] TABULA-typology approach for building stock energy assessment, EPISCOPE, <http://episcope.eu>,
737 [Online; accessed 10-July-2019].
- 738 [20] Commercial Reference Buildings, US Department of Energy, (DOE), <https://energy.gov/eere/buildings/commercial-reference-buildings>, [Online; accessed 10-July-2019].
- 740 [21] ENTRANZE-ENforce the TRAnSition to Nearly Zero energy buildings, Intelligent Energy Europe pro-
741 gramme, <http://www.entranze.eu>, [Online; accessed 10-July-2019].
- 742 [22] The Buildings Performance Institute Europe (BPIE), European Union, <http://bpie.eu>, [Online; ac-
743 cessed 10-July-2019].
- 744 [23] X. Li, R. Yao, M. Liu, V. Costanzo, W. Yu, W. Wang, A. Short, B. Li, Developing urban residential
745 reference buildings using clustering analysis of satellite images, *Energy and Buildings* 169 (2018) 417–
746 429.
- 747 [24] J. Torriti, A review of time use models of residential electricity demand, *Renewable and Sustainable*
748 *Energy Reviews* 37 (2014) 265–272.
- 749

- 750 [25] M. Peña, F. Biscarri, J. I. Guerrero, I. Monedero, C. León, Rule-based system to detect energy efficiency
751 anomalies in smart buildings, a data mining approach, *Expert Systems with Applications* 56 (2016)
752 242–255.
- 753 [26] S. Singh, A. Yassine, Big data mining of energy time series for behavioral analytics and energy con-
754 sumption forecasting, *Energies* 11 (2) (2018) 452.
- 755 [27] K. Amasyali, N. M. El-Gohary, A review of data-driven building energy consumption prediction studies,
756 *Renewable and Sustainable Energy Reviews* 81 (2018) 1192–1205.
- 757 [28] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches
758 for prediction and classification of building energy consumption, *Renewable and Sustainable Energy*
759 *Reviews* 82 (2018) 1027–1047.
- 760 [29] É. Mata, A. Sasic Kalagasidis, F. Johnsson, Building-stock aggregation through archetype buildings:
761 France, Germany, Spain and the UK, *Building and Environment* 81 (2014) 270–282.
- 762 [30] Building Performance Database by Department of Energy, <https://bpd.lbl.gov/>, [Online; accessed
763 10-July-2019].
- 764 [31] Censushub - European Statistical System (ESS), <https://ec.europa.eu/CensusHub2>, [Online; ac-
765 cessed 10-July-2019].
- 766 [32] Data hub for energy performance of buildings, buildings performance institute europe (BPIE), [https://ec.europa.eu/energy/en/topics/energy-efficiency/energy-performance-of-buildings/
767 eubuildings](https://ec.europa.eu/energy/en/topics/energy-efficiency/energy-performance-of-buildings/eubuildings), [Online; accessed 10-July-2019] (2017).
- 768 [33] Assessment and improvement of the EPBD impact (for new buildings and building renova-
769 tion) (ASIEPI), Intelligent Energy Europe- European Commission, [https://ec.europa.eu/energy/
770 intelligent/projects/en/projects/asiapi](https://ec.europa.eu/energy/intelligent/projects/en/projects/asiapi), [Online; accessed 10-July-2019].
- 771 [34] Energy Performance Certificates across europe from design to imple-
772 mentation, BPIE, [http://www.buildup.eu/en/practices/publications/
773 energy-performance-certificates-across-europe-design-implementation](http://www.buildup.eu/en/practices/publications/energy-performance-certificates-across-europe-design-implementation), [Online; accessed
774 10-July-2019] (2010).
- 775 [35] E. G. Dascalaki, K. Droutsas, A. G. Gaglia, S. Kontoyiannidis, C. A. Balaras, Data collection and
776 analysis of the building stock and its energy performance example for hellenic buildings, *Energy and*
777 *Buildings* 42 (8) (2010) 1231–1237.
- 778 [36] J. Keirstead, M. Jennings, A. Sivakumar, A review of urban energy system models: Approaches,
779 challenges and opportunities, *Renewable and Sustainable Energy Reviews* 16 (6) (2012) 3847–3866.
- 780 [37] S. Shnapp, Robust building data: A driver for policy development, in: *World Sustainable Energy Days*
781 *Next 2014*, Springer, 2015, pp. 3–10.
- 782 [38] I. Ballarini, S. P. Corgnati, V. Corrado, N. Talà, Improving energy modeling of large building stock
783 through the development of archetype buildings, in: *12th Conference of the International Building*
784 *Performance Simulation*. Australia, Sydney, 2011, pp. 14–16.
- 785 [39] A. Schaefer, E. Ghisi, Method for obtaining reference buildings, *Energy and Buildings* 128 (2016)
786 660–672.
- 787 [40] O. Neu, B. Sherlock, S. Oxizidis, D. Flynn, D. Finn, Developing building archetypes for electrical load
788 shifting assessment: Analysis of irish residential stock (2014).
- 789 [41] J. Kragh, K. B. Wittchen, Development of two danish building typologies for residential buildings,
790 *Energy and Buildings* 68 (2014) 79–86.
- 791 [42] M. H. Kristensen, R. E. Hedegaard, S. Petersen, Hierarchical calibration of archetypes for urban building
792 energy modeling, *Energy and Buildings* 175 (2018) 219–234.
- 793 [43] E. G. Dascalaki, K. G. Droutsas, C. A. Balaras, S. Kontoyiannidis, Building typologies as a tool for
794 assessing the energy performance of residential buildings—a case study for the hellenic building stock,
795 *Energy and Buildings* 43 (12) (2011) 3400–3409.
- 796 [44] T. Csoknyai, S. Hrabovszky-Horváth, Z. Georgiev, M. Jovanovic-Popovic, B. Stankovic, O. Villatoro,
797 G. Szendrő, Building stock characteristics and energy performance of residential buildings in eastern-
798 european countries, *Energy and Buildings* 132 (2016) 39–52.
- 799 [45] L. Filogamo, G. Peri, G. Rizzo, A. Giaccione, On the classification of large residential buildings stocks
800

- 801 by sample typologies for energy planning purposes, *Applied Energy* 135 (2014) 825–835.
- 802 [46] Residential Prototype Building Models by U.S. Department of Energy’s Building Energy Codes Pro-
803 gram), https://www.energycodes.gov/development/residential/iecc_models, [Online; accessed
804 10-July-2019] (2012).
- 805 [47] T. Loga, B. Stein, N. Diefenbach, Tabula building typologies in 20 european countriesmaking energy-
806 related features of residential building stocks comparable, *Energy and Buildings* 132 (2016) 4–12.
- 807 [48] TABULA, Typology approach for building stock energy assessment. main results of the tabula project
808 (2012).
- 809 [49] I. Theodoridou, A. M. Papadopoulos, M. Hegger, Statistical analysis of the greek residential building
810 stock, *Energy and Buildings* 43 (9) (2011) 2422–2428.
- 811 [50] U. Ali, C. Buccella, C. Cecati, Households electricity consumption analysis with data mining techniques,
812 in: *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, 2016, pp. 3966–
813 3971. doi:10.1109/IECON.2016.7793118.
- 814 [51] C. Fan, F. Xiao, Y. Zhao, A short-term building cooling load prediction method using deep learning
815 algorithms, *Applied Energy* 195 (2017) 222–233.
- 816 [52] C. Zhang, L. Cao, A. Romagnoli, On the feature engineering of building energy data mining, *Sustainable
817 Cities and Society* 39 (2018) 508–518.
- 818 [53] D.-S. Kapetanakis, E. Mangina, D. P. Finn, Input variable selection for thermal load predictive models
819 of commercial buildings, *Energy and Buildings* 137 (2017) 13–26.
- 820 [54] J. Egan, D. Finn, P. H. D. Soares, V. A. R. Baumann, R. Aghamolaei, P. Beagon, O. Neu, F. Pallonetto,
821 J. O’Donnell, Definition of a useful minimal-set of accurately-specified input data for building energy
822 performance simulation, *Energy and Buildings* 165 (2018) 172–183.
- 823 [55] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, Lof: Identifying density-based local outliers, in:
824 *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ACM, 2000,
825 pp. 93–104. doi:10.1145/342009.335388.
- 826 [56] U. Ali, M. H. Shamsi, F. Alshehri, E. Mangina, J. O’Donnell, Comparative analysis of machine learning
827 algorithms for building archetypes development inurban building energy modeling (2018).
- 828 [57] G. Tardioli, R. Kerrigan, M. Oates, J. O’Donnell, D. P. Finn, Identification of representative build-
829 ings and building groups in urban datasets using a novel pre-processing, classification, clustering and
830 predictive modelling approach, *Building and Environment* 140 (2018) 90–106.
- 831 [58] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,
832 *Journal of computational and applied mathematics* 20 (1987) 53–65.
- 833 [59] D. L. Davies, D. W. Bouldin, A cluster separation measure, *IEEE transactions on pattern analysis and
834 machine intelligence PAMI-1* (2) (1979) 224–227.
- 835 [60] T. F. Liao, Measuring and analyzing class inequality with the gini index informed by model-based
836 clustering, *Sociological Methodology* 36 (1) (2006) 201–224.
- 837 [61] G. Medina Benejam, Bottom-up characterisation of the spanish building stock–archetype buildings and
838 energy demand (2011).
- 839 [62] R. Arababadi, Energy use in the eu building stock-case study: Uk, Master’s thesis, Linkping University
840 (2012).
- 841 [63] Census of Population 2016 - Profile 1 Housing in Ireland by Central Statistics Office, [https://www.cso.
842 ie/en/releasesandpublications/ep/p-cp1hii/cp1hii/hs/](https://www.cso.ie/en/releasesandpublications/ep/p-cp1hii/cp1hii/hs/), [Online; accessed 10-July-2019] (2016).
- 843 [64] Y. Zhang, I. Korolija, Performing complex parametric simulations with jeplus, in: *SET2010-9th Inter-
844 national Conference on Sustainable Energy Technologies*, 2010, pp. 24–27.
- 845 [65] V. Corrado, H. E. Mechri, Uncertainty and sensitivity analysis for building energy rating, *Journal of
846 building physics* 33 (2) (2009) 125–156.
- 847 [66] O. Neu, S. Oxizidis, D. Flynn, D. Finn, Utilising time of use surveys to predict domestic hot water
848 consumption and heat demand profiles of residential building stocks, *British Journal of Environment
849 and Climate Change* 6 (2) (2016) 77–89.
- 850 [67] U. Ali, M. H. Shamsi, C. Hoare, J. O’Donnell, Gis-based residential building energy modeling at
851 district scale, in: *BSO 2018: 4th IBPSA-England Conference on Building Simulation and Optimization*,

852 Cambridge, United Kingdom, 11-12 September 2018, International Building Performance Simulation
853 Association, 2018, pp. 153–160.