2018 Building Performance Modeling Conference and
SimBuild co-organized by ASHRAE and IBPSA-USA
Chicago, IL
September 26-28, 2018

COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMSFOR BUILDING ARCHETYPES DEVELOPMENT IN URBAN BUILDING ENERGY MODELING

Usman Ali[1], Mohammad Haris Shamsi[1], Fawaz Alshehri[1], Eleni Mangina[2] and James O'Donnell[1]
[1]School of Mechanical & Materials Engineering, Energy Institute
University College Dublin (UCD), Ireland
[2]School of Computer Science and Informatics, University College Dublin (UCD), Ireland

ABSTRACT

The most common approach for urban building energy modeling (UBEM) involves segmenting a building stock into archetypes. Development Building archetypes for urban scale is a complex task and requires a lot of extensive data. The archetype development methodology proposed in this paper uses unsupervised machine learning approaches to identify similar clusters of buildings based on building specific features. The archetype development process considers four crucial processes of machine learning: data preprocessing, feature selection, clustering algorithm adaptation and results validation. The four different clustering algorithms investigated in this study are K-Mean, Hierarchical, Density-based, K-Medoids. All the algorithms are applied on Irish Energy Performance Certificate (EPC) that consist of 203 features. The obtained results are then used to compare and analyze the chosen algorithms with respect to performance, quality and clus-ter instances. The K-mean algorithm preforms the best in terms of cluster formation.

INTRODUCTION

The most common approach for bottom-up UBEM involves segmenting a building stock into archetypes (Sokol, Davila, and Reinhart 2017). Buildings possessing similar characteristics are usually grouped together representing a large building stock and are termed as archetypes (Galante et al. 2012; Famuyibo, Duffy, and Strachan 2012). Therefore, an archetype is a virtual building that represents the number of buildings sharing similar characteristics in the stock (Sousa Monteiro et al. 2015). Because of the underlying difficulties in gathering detailed information at an urban scale, the archetypes approach has become popular in urban energy modeling as the available information can be used to model similar buildings. In contrast, for individual building models, detailed information is usually collected from surveys and architectural and mechanical designs (Sokol, Davila, and Reinhart 2017).

The building stock can be classified into three categories namely building typologies, building archetypes and ref-erence buildings. The building typologies concept cat-egorizes the buildings into groups by the similarity of their use such as residential, office, school etc. The building archetypes concept is most commonly used in energy modeling at the urban scale. Reference buildings concept is used by the European Union Energy Performance of Buildings Directive (EPBD). This Directive 2010/31/EU (Recast 2010) introduces the concept of cost-optimal frameworks which are used for calculating cost-optimal levels of minimum energy performance requirements for buildings and building elements. According to the Commission Delegated Regulation No. 244/ 2012 (EU-Commission et al. 2012), member states are required to define reference buildings that should represent the average building stock in each member state (Ballarini, Corgnati, and Corrado 2014).

Several projects (de Vasconcelos et al. 2015), both at the EU and international level, are being developed in order to define the building stock, such as TABULA (Loga, Diefenbach, and Stein 2012), ASIEPI (Intelligent Energy Europe 2018), BPIE (European Union 2018) and DOE (US Department of Energy 2018). The first major project called Typology Approach for Building Stock Energy Assessment (TABULA) was developed to construct a European database of building typologies. Similarly, the primary goal of ASsessment and Improvement of the EPBD Impact for new buildings and building renovation (ASIEPI), was to provide support on EPBD related aspects that may present potential problems when implementing the EPBD in the Member States and the European Commission. Similarly, the Buildings Performance Institute Europe (BPIE) has undertaken an extensive survey across Europe to improve the energy performance of buildings. The BPIE also launched the Data Hub (The Buildings Performance Institute Europe 2018) portal for gathering statistical data on the comprehensive snapshot of the building stock characteristics in European Union. The United States, Department of Energy (DOE), devel-oped commercial reference buildings, formerly known as commercial building benchmark models. The developed building types represent approximately 70% of the commercial buildings in the United States.

Famuyibo et al. proposed a detailed statistical analysis method for archetypes development which allows for a detailed representation of the overall building stock as compared to the traditional qualitative techniques (Famuyibo, Duffy, and Strachan 2012). The author used multi-linear regression analysis and descriptive statis-tics for the identification of archetypes. The developed archetypes were representative of 65% of the population Irish housing stock. Similarly, Schaefer et al. used clus-ter analysis approach to obtain reference buildings but the case study only covers the low-income housing stock in Florianpolis, Southern Brazil (Schaefer and Ghisi 2016). Lara et al. used clustering and regression analysis approach to identify the most suitable parameters in the classification of a large sample of existing buildings (Arambula Lara et al. 2014).

The major issue with all previous studies is the absence of comparison of the clustering techniques. There should be an opportunity to compare and evaluate different cluster-ing techniques to achieve a better result. Another draw-back is that all the approaches are tested on the specific area or construction period, and there should be a gener-alized statistical analysis method that can be used in any scenario.

This paper proposes a generalized approach for building archetypes development in urban energy modeling. To achieve better results, comparative analysis of differ-ent clustering algorithms is performed to identify the best method.

METHODOLOGY

The development of archetypes using machine learning techniques requires following steps as demonstrated in Figure 1. The methodology process follows the standardized procedure of process analysis in data science, which initiates with data collection followed by pre-processing, elimination of outliers, implementation of analysis algorithms and terminates with results validation.

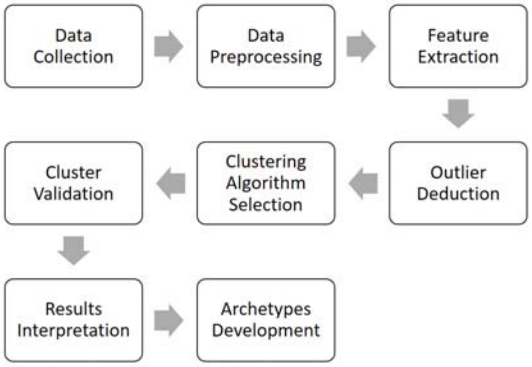

Figure 1: Methodology for the building archetypes devel-opment using machine learning algorithms

Data collection

Generally, building stocks are categorized into residential (R), also called domestic or household sector, and non-residential (NR) buildings, known as commercial sector. Residential buildings comprise of all types of houses including detached, semi-detached, terraced houses, houses built in a row, etc while non-residential buildings comprise industrial, commercial, educational and health care installations. Typically, the information for characterizing building stocks is gathered through two approaches such as national census databases and statistical survey such as Buildings Performance Database (BPD) and Commercial Building Resource Database (Mata, Sasic Kalagasidis, and Johnsson 2014). There are other additional methods for gathering building data such as individual billing data, sub-metering, energy certificates (Majcen, Itard, and Visscher 2013) or through geographical information systems (Mastrucci et al. 2014; Torabi Moghadam, Mutani, and Lombardi 2016).

Data pre-processing

The data obtained through measurement systems or through extensive surveys is often incomplete and lacks certain important variables. Furthermore, the data needs to be checked for errors, noises or outliers. Hence, it is crucial to pre-process the data before the implementation of machine learning algorithms. Data pre-processing al-lows to achieve accurate prediction results. Some of the major pre-processing techniques are data cleaning, inte-gration, transformation, reduction and discretization (Ali, Buccella, and Cecati 2016).

Feature selection

The next essential step in model clustering is the input fea-ture selection. Feature selection method has been widely used to obtain most representative and useful variables from the data. The basic goal of feature selection is to select the most appropriate inputs for the model because historical data often possesses a lot of irrelevant or redun-dant variables. One of the main advantages of feature selection is the reduced dimensionality of the model in-puts that significantly reduces the computational load and increases the model accuracy. Feature selection is usu-ally performed using engineering and statistical methods. The engineering methods are mainly based on experts' in-terpretations and existing practices in the literature. The statistical methods use different statistical or data mining methods such as mean and standard deviation, regression techniques, genetic algorithms etc (Fan, Xiao, and Zhao 2017). In this paper, engineering methods are used for feature selection. The goal of archetypes development is further used for urban energy modeling. A number of studies have identified the minimal number of crucial fea-tures required to facilitate the energy modeling process

(Famuyibo, Duffy, and Strachan 2012; Egan et al. 2018).

Outlier detection

Outlier detection is a crucial task in machine learning and aims to find the observation that possesses noise, exceptional dis-similar information and is inconsistent with the majority of the data. Handling noisy data is important before implementing clustering algorithms on a large dataset. This technique could be a part of data pre-processing. The outlier detection techniques are distance-based, density-based and Local Outlier Factor (LOF). In this paper, LOF algorithm is used for detecting the outliers. LOF is an outlier algorithm that was proposed by Breunig et al. for finding the outlier data points utiliz-ing the nearest neighbors to calculate the local deviation of each given data point (Breunig et al. 2000). LOF of an object p is the average of the ratio of local reachability density of p and those of $p's$ nearest neighbors (NN). LOF can be computed using Equation (1).

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

where, p and p' are two data points.

Clustering algorithms

Clustering is a technique of assigning a set of objects to the same group (called a cluster), so that the objects in a particular cluster have similar values to each other than to those in other clusters. In this paper, following four clus-tering techniques are tested to investigate the appropriate technique for archetypes development.

K-means clustering

The K-means is the most common unsupervised partitional classification algorithm to solve the well-known clustering problem (MacQueen et al. 1967). Each clus-ter is represented by the mean of the cluster. The aim of the k-means algorithm is to divide the observations into k clusters in which each observation belongs to the respec-tive cluster (center point). The objective is to minimize the sum of distances of the points to their respective cen-troid. The most common definition is with Euclidean dis-tance, minimizing the Sum of Squared Error (SSE) func-tion. The objective function is given in Equation (2).

$$C = \sum_{k=1}^{K} \sum_{i \in c_k} \|x_i - m_k\|^2 \text{ Euclidean distance (2)}$$

where c_k is the mean of the n data points in cluster C_i. The simplicity and scalability of this approach is the main advantage of K-mean clustering technique when compared to other algorithms. K-means has major limitations when clusters are of different sizes, densities and when the data contains outliers.

K-Medoids clustering

K-medoids clustering is also a partition based algorithm in which each cluster is represented by one of the data points located near the center of the cluster. The K-medoids is a variant of K-means technique that is more robust to noises and outliers, and uses mediods to repre-sent the cluster rather than centroids. Partitioning Around Medoids (PAM) is a one of the representative K-medoids clustering method (Kaufman and Rousseeuw 2009).

Hierarchical clustering

Hierarchical clustering is a method of cluster analy-sis which seeks to build a hierarchy of clusters (John-son 1967). Hierarchical clustering falls into two types, namely, agglomerative and divisive. The agglomerative approach is a bottom up approach which starts with each individual cluster and at each step, pairs and merges with the closest clusters as one moves up the hierarchy using a predefined linkage method. The divisive approach is top-down approach that starts with one cluster, and at each step, splits the cluster recursively as one moves down the hierarchy. The benefit of hierarchical clustering is it's easy implementation and provision of better results in some cases. The major limitation is the computational complexity involved in time and space. In this paper, an agglomerative algorithm is used for hierarchical test clustering because of it's ability to directly define the similarity between different clusters. Generally, there are three ways to update the distances in the agglomerative algorithm such as single, complete or average linkages. Different linkages have different partitioning approaches, so the type of linkage is selected by analyzing the type of data to be clustered.

Density-based clustering

Density-based clustering algorithm, for instance, Density-based spatial clustering of applications with noise (DBSCAN), (Ester et al. 1996) partitions the points into dense regions separated by less dense regions. In DBSCAN, the density at any point p is the number of points within a circle of radius Eps. The dense region represents a circle of radius Eps that contains at least $MinPts$ points. The density based clustering approach can discover arbitrary-shapes of clusters with varying size and the technique is insensitive to noise and outliers in the data. The major limitations include the complexities involved in computation and higher sensitivity to input features.

Cluster validation

The result validation process is the last and the most im-portant step to identify whether the implemented algo-rithms are practically relevant. Certain criteria and valid-ity indexes establish the relevancy of each implemented algorithm. To measure the validity of clustering results, the internal validity indices are used that calculate the properties of resulting clusters, such as compactness, separation and roundness. The most common validity indices are silhouette index (Rousseeuw 1987), davies bouldin index, gini index and cophenetic correlation coefficient.

Table 1: Dublin Energy Performance Certificate (EPC) dwelling types for clustering analysis

Dwelling Type	Buildings	Percentage
Semi-detached house	48863	24.15%
Mid-terrace house	40257	19.90%
Mid-floor apartment	31142	15.39%
End of terrace house	19767	9.77%
Ground-floor apartment	19124	9.45%
Top-floor apartment	18735	9.26%
Detached house	15553	7.69%
House	3344	1.65%
Maisonette	3086	1.53%
Apartment	2309	1.14%
Basement Dwelling	110	0.05%

The silhouette of a cluster value is a measure of the number of objects that lie well inside the own cluster and which do not. The calculation is based on the silhouette width of their cluster objects. Mathematically, the silhouette width for each object j can be represented by Equation 3.

$$S(j) = \frac{b(j) - a(j)}{maxa(j), b(j)} \tag{3}$$

where $a(j)$ is the average dissimilarity between j and all other objects in the cluster, and $b(j)$ is the minimum of the average dissimilarity between j and objects in other clusters. The silhouette index (SI) is a normalized index and a value close to 1 is always good for clustering.

The DaviesBouldin Index (DBI) (Davies and Bouldin 1979) is the ratio within cluster distances to between-cluster separation. Therefore, if the clusters are compact and well separated, the value of the DBI is small which is ideal for clustering. The DBI is defined as in Equation 4:

$$V_{DB} = \frac{1}{k} \sum_{i=1}^{k} R_i \tag{4}$$

where k is the number of clusters and R_i is defined as in Equation 5:

$$R_i = \max_{i \neq j} R_{ij} \tag{5}$$

where $R_i j$ is the similarity measure between clusters C_i and C_j, and is defined as:

$$R_{ij} = \frac{S_i + S_j}{D_{ij}} \tag{6}$$

The item distribution measure gives the idea about the size of clusters This performance measure of a cluster is evaluated using two indexes, namely, sum of squares and Gini coefficient. The Gini coefficient, also known as the Gini Index (GI), is a measure of statistical dispersion or cluster competence. A higher GI indicates an unequal distribution, while a lower GI suggests an equal distribution.

Cophenetic correlation coefficient (CPCC) can be used to evaluate the efficiency of hierarchical clustering techniques that utilize different linkage methods, for instance single, complete or average linkages (Saraçli, Doğan, and Doğan 2013). A high value of CPCC indicates good hierarchical clustering techniques.

Results interpretation

This step describes the interpretation of clustering results on the basis of calculated validity indices as stated in the previous step. Different clustering algorithms are used for different purposes. Also a few algorithms work better on specific data as compare to other. Selection of algorithms and the number of clusters is a difficult task and sometimes a trade-off between the different validity indices is required to achieve better results.

Archetypes development

The last step in the process involves the development of building archetypes on the basis of developed clusters. Each cluster represents one archetype of building and all the variables selected in the feature selection phase represent the characteristics of that building.

RESULTS AND DISCUSSION

The main objective of the paper is to develop the building archetypes that represent an entire urban area. The methodology presented above is applied to the publicly available Irish Energy Performance Certificate (EPC) data published by the Sustainable Energy Authority of Ireland (SEAI). The EPC data is used to evaluate each building's energy performance. The certificate rates the energy performance of a building in terms of primary energy consumption (kWh/m^2/year) and varies on a scale of A to G. An A-rated building represents a building with the highest energy efficiency and will tend to have the lowest energy consumption and subsequent lower CO_2 emissions. On the other hand, a G-rated building represents a building with the lowest energy efficiency. The EPC is calculated with Dwelling Energy Assessment Procedure (DEAP) software, that is Ireland's official method for calculating the building energy rating of new and existing buildings. The EPC data contains more than 600,000 Irish buildings' data with 203 variables including building physics, energy, and CO_2 information. The city of Dublin, which contains more than 200,000 buildings, is chosen for the archetype development using clustering al-

Table 2: Number of clusters with the best validity values of k-mean and agglomerative clustering analysis

Dwelling Types	**K-Mean**					**Agglomerative**		
	k	**AD**	**DBI**	**GI**	**SI**	**k**	**GI**	**CPCC**
Mid-floor apartment	2	451	0.75	1.00	0.61	7	1.00	0.88
Top-floor apartment	2	725	0.97	1.00	0.42	8	0.99	0.85
Mid-terrace house	2	445	0.70	1.00	0.60	10	0.99	0.96
Semi-detached house	4	226	0.82	0.98	0.39	10	0.91	0.99
Detached house	3	175	0.25	0.88	0.68	6	0.81	0.93
Maisonette	10	326	0.82	0.97	0.42	10	0.93	0.87
Ground-floor apartment	6	262	0.98	1.00	0.31	10	0.99	0.76
House	4	535	0.51	0.89	0.34	4	0.79	0.87
Apartment	2	2035	0.46	1.00	0.78	10	0.97	0.93
End of terrace house	7	110	0.78	0.99	0.40	10	0.97	0.84
Basement Dwelling	7	104	0.69	0.90	0.39	9	0.83	0.89

Table 3: Number of clusters with the best validity index values of k-medoids and density-based clustering analysis

Dwelling Types	**K-Medoids**					**DBSCAN**				
	k	**AD**	**DBI**	**GI**	**SI**	**Ep**	**Mp**	**C**	**SI**	**GI**
Mid-floor apartment	2	530	0.75	1.00	0.62	0.36	10	5	-0.03	1.00
Top-floor apartment	2	1014	1.09	1.00	0.37	0.46	8	2	0.06	1.00
Mid-terrace house	2	749	1.18	1.00	0.59	0.2	8	1	1.00	1.00
Semi-detached house	2	1156	0.69	1.00	0.42	0.2	3	1	1.00	1.00
Detached house	2	969	0.21	0.93	0.63	0.2	2	1	1.00	1.00
Maisonette	10	532	0.98	0.97	0.38	0.2	4	1	1.00	1.00
Ground-floor apartment	8	373	1.67	1.00	0.22	0.2	8	1	1.00	1.00
House	2	2602	0.57	0.93	0.51	0.2	2	1	1.00	1.00
Apartment	2	3648	0.54	1.00	0.78	0.4	4	10	0.03	0.98
End of terrace house	2	649	0.63	1.00	0.50	0.2	7	1	1.00	1.00
Basement Dwelling	2	621	0.74	0.97	0.48	0.2	2	1	1.00	1.00

gorithms. The data is further divided into the subset of 11 dwelling types as shown in Table 1. The clustering algo-rithms take huge computational time to classify the whole city data. Hence, the comparison of the clustering algo-rithms is performed only for a specific region, Dublin 1, in the district that represents 6658 buildings of the entire Dublin city.

After the data collection, the next step, i.e, data preprocessing is applied to clean the data such as replacing the missing values with average, removing useless variables or less frequent values using standard deviation threshold. Clustering is mostly applied on numerical val-ues so all the nominal values are converted to numerical. The EPC data contains 203 variables which are sorted out to remove all the irrelevant variables for archetypes devel-opment. There are two methods discussed in the previous section for feature selection. In this paper, existing lit-erature is considered for feature selection. By using the method devised by Famuyibo et al., 14 variables are iden-tified as relevant and would influence the building's en-ergy performance (Famuyibo, Duffy, and Strachan 2012). Some of these these features include building u-values, building areas, primary fuel source, etc.

The LOF algorithm is used for outlier removal from the EPC data. LOF is based on the distance function to mea-sure the density of objects amongst each other. The Eu-clidean distance measure is used with a lower bound of 10 *MinPts* and an upper bound of 20 *MinPts*.

K-mean clustering

The K-mean algorithm using Euclidean distance is applied to the Dublin 1 city dataset and the best number of classes are chosen for each dwelling type. Best values for each dwelling type are shown in Table 2. The result shows the best value of K-mean validity indices, for instance, Average within Distance (AD), SI, GI and DBI. The max-imum and minimum number of clusters are found to be 10 and 2 and the total number of clusters that represent the archetypes in Dublin 1, is found to be 49.

Table 4: K-mean number of clusters analysis of Dublin City for archetypes development

Dwelling Types	k	Number of clusters and items	Total
Mid-floor apartment	2	Cluster 0: 2206, Cluster 1: 28540	30746
Top-floor apartment	2	Cluster 0: 6217, Cluster 1: 12264	18481
Mid-terrace house	2	Cluster 0: 32313, Cluster 1: 7572	39885
Semi-detached house	4	Cluster 0: 19704, Cluster 1: 1492, Cluster 2: 19257, Cluster 3: 8021	48474
Detached house	3	Cluster 0: 7803, Cluster 1: 2549, Cluster 2: 827, Cluster 3: 4230	15409
Maisonette	10	Cluster 0: 570, Cluster 1: 281, Cluster 2: 214, Cluster 3: 199, Cluster 4: 70, Cluster 5: 860, Cluster 6: 61, Cluster 7: 261, Cluster 8: 398, Cluster 9: 113	3027
Ground-floor apartment	6	Cluster 0: 5245, Cluster 1: 3104, Cluster 2: 4642, Cluster 3: 4737, Cluster 4: 951, Cluster 5: 248	18927
House	4	Cluster 0: 320, Cluster 1: 1927, Cluster 2: 40, Cluster 3: 977	3264
Apartment	2	Cluster 0: 20, Cluster 1: 2198	2218
End of terrace house	7	Cluster 0: 1611, Cluster 1: 1918, Cluster 2: 5078, Cluster 3: 1228, Cluster 4: 8675, Cluster 5: 816, Cluster 6: 185	19511
Basement Dwelling	7	Cluster 0: 8, Cluster 1: 31, Cluster 2: 8, Cluster 3: 4, Cluster 4: 27, Cluster 5: 7, Cluster 6: 22	107

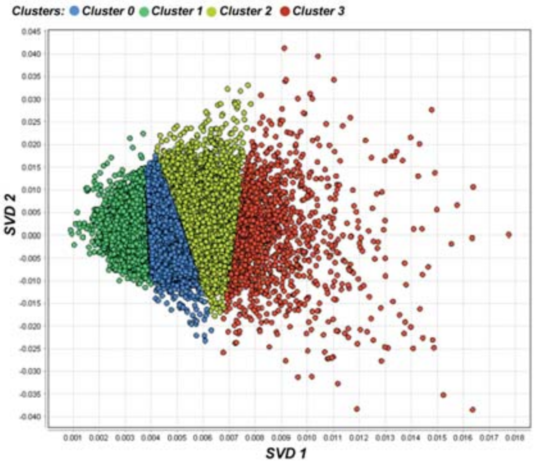

Figure 2: 2-Dimensions singular value decomposition scatter plot of Dublin city semi-detached house for k-mean cluster analysis

Hierarchical clustering

The agglomerative algorithm is used to implement the hierarchical clustering utilizing the best linkage method. To examine the different linkage and distance methods for archetypes development, an explanatory analysis is per-formed on the dataset using validity indices such as CPCC and GI. The best average linkage method is the one which creates minimum 4 and maximum 10 number of clusters as shown in Table 2. Total 94 clusters are created which are quite large in number as compared to the ones created using the other cluster algorithms. The reason behind the large number of clusters is due to the complexity inhibited by the input variables which possess different characteris-tics.

K-medoids clustering

The K-medoids algorithm using euclidean distance is applied on the Dublin 1 city dataset and the best number of classes for each dwelling type are selected. The best value for each dwelling type is shown in Table 3. The result in-dicates the best value of k-medoids validity indices, for instance, AD, SI, GI and DBI. The minimum 2 clusters and maximum 10 clusters are formed and the total num-ber of clusters that represents the archetypes in Dublin 1, are found to be 48.

Density-based clustering

DBSCAN approach is sensitive to the parameter epsilon (Ep) and minPts (Mp), which determine the optimal number of clusters and more validity indexes such as SI and GI are used. The best value for each dwelling type is shown in Table 3. The result shows that this algorithm is not suit-able for the EPC dataset as DBSCAN best works with a large dataset while some of the dwelling types possess a small dataset. As such, only 1 cluster is developed with the best validity index.

After the analysis of all the aforementioned clustering algorithms, K-mean is found to be the most suitable (for this research) and reliable. K-mean approach performs bet-ter in terms of the consistent distribution of the k classes and certroid values as compared to the other algorithms. The entire Dublin city is then analyzed using the K-mean approach. Table 4 represents the number of clusters and items distribution for the K-mean approach. For example, clusters of the semi-detached house using the value of k as 4 generated a total number of 48474 clusters items in which the cluster 0 to 3 contain items 19704, 1492, 19257 and 8021 respectively. Singular Value Decomposition (SVD) is used for better understanding of K-mean cluttered data by focusing on the number of important dimensions that are shown in Figure 2. This observation relates to the fact that there are four types of archetypes in a semi-detached house and the results show that all the developed clusters are compacted and separated to each other. Similarly, an identical approach is applied on all other dwelling types.

Hence, a total number of 49 archetypes are identified that represent the entire city of Dublin. The centroid values of one cluster represent the characteristics of each archetype. Each archetype has unique values of selected features such as construction material, glazing (U-Values), heating and cooling system etc. For instance, 4 clusters of the semi-detached dwelling type are identified. The centroid of each cluster represents the characteristic of that particular archetype. The windows U-values of 4 buildings archetypes are 2.65, 2.73, 3.02 and 2.85 (W/m^{2}K). Similarly, average windows areas associated with each archetype are 31.81, 23.29, 17.36 and 45.33 (m^2).

The research analysis conducted in this paper identifies 36 building archetypes from 11 dwelling types based on key variables from existing literature that represent more than 200,000 buildings. Each archetype building represent a cluster having common building characteristics. The proposed methodology is flexible enough to include more key variables and be applied to a different dataset.

CONCLUSION

Accurate characterization of the existing building stock has become essential to ensure efficient implementation of the energy efficiency measures. The research con-ducted in this paper proposes a methodology to develop the archetypes of different dwelling types which will al-low for an extended and a detailed analysis of the energy performance of buildings. These archetypes could be fur-ther used as benchmarks or reference buildings to evaluate the energy savings and efficiency strategies. The proposed generalized archetype development methodology consists of several machine learning steps. Furthermore, a compar-ative study of unsupervised machine learning algorithms for archetypes development in UBEM is conducted. The devised methodology is applied to the Dublin city consisting of more than 200,000 buildings. The analysis resulted in the identification of 49 archetypes that represent the building stock of the entire Dublin city. K-mean algo-rithm performs better than the other ones when compared in terms of the optimal value of the indices considered. The archetypes of buildings obtained can be used as a guideline for construction of new buildings and standard assessment methodologies to improve the building performance on a large scale. Furthermore, the developed archetypes would aid in the implementation of retrofit strategies on existing buildings at a district scale.

The identified clusters will further aid the urban planners in creating retrofit strategies to improve the building energy performance at a large scale.

Future work could involve further testing of the developed archetypes through energy simulation modeling such as using EnergyPlus.

ACKNOWLEDGMENT

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

REFERENCES

Ali, Usman, Concettina Buccella, and Carlo Cecati. 2016. “Households electricity consumption analysis with data mining techniques.” *Industrial Electron-ics Society, IECON 2016-42nd Annual Conference of the IEEE.* IEEE, 3966–3971.

Arambula Lara, Rigoberto, Francesca Cappelletti, Piercarlo Romagnoni, and Andrea Gasparella. 2014. Se-lection of Representative Buildings through Prelimi-nary Cluster Analysis.

Ballarini, Ilaria, Stefano Paolo Corgnati, and Vincenzo Corrado. 2014. “Use of reference buildings to assess the energy saving potentials of the residential building stock: The experience of TABULA project.” *En-ergy Policy* 68:273–284.

Breunig, Markus M, Hans-Peter Kriegel, Raymond T Ng, and J¨org Sander. 2000. “LOF: identifying density-based local outliers.” *ACM sigmod record*, Volume 29. ACM, 93–104.

Davies, David L, and Donald W Bouldin. 1979. “A clus-ter separation measure.” *IEEE transactions on pat-tern analysis and machine intelligence*, no. 2:224–227.

de Vasconcelos, Ana Brand˜ao, Manuel Duarte Pinheiro, Armando Manso, and Ant´onio Cabac¸o. 2015. “A Portuguese approach to define reference buildings for cost-optimal methodologies.” *Applied Energy* 140:316–328.

Egan, James, Donal Finn, Pedro Henrique Deogene Soares, Victor Andreas Rocha Baumann, Reihaneh Aghamolaei, Paul Beagon, Olivier Neu, Fabiano Pallonetto, and James ODonnell. 2018. “Definition of a useful minimal-set of accurately-specified input data for Building Energy Performance Simulation.” *En-ergy and Buildings* 165:172–183.

Ester, Martin, Hans-Peter Kriegel, J¨org Sander, Xiaowei Xu, et al. 1996. “A density-based algorithm for discovering clusters in large spatial databases with noise.” *Kdd*, Volume 96. 226–231.

EU-Commission, et al. 2012. Commission Delegated Regulation (EU) No 244/2012 of 16 January 2012 supplementing Directive 2010/31.

European Union, EU. 2018. The Buildings Performance Institute Europe (BPIE).

Famuyibo, Adesoji Albert, Aidan Duffy, and Paul Strachan. 2012. “Developing archetypes for domestic dwellingsan Irish case study.” *Energy and Buildings* 50:150–157.

Fan, Cheng, Fu Xiao, and Yang Zhao. 2017. “A short-term building cooling load prediction method using deep learning algorithms.” *Applied Energy* 195:222–233.

Galante, Annalisa, Marco Torri, et al. 2012. “A methodology for the energy performance classification of residential building stock on an urban scale.” *Energy and Buildings* 48:211–219.

Intelligent Energy Europe, European Commission. 2018. Assessment and improvement of the EPBD Impact (for new buildings and building renovation) (ASIEPI).

Johnson, Stephen C. 1967. “Hierarchical clustering schemes.” *Psychometrika* 32 (3): 241–254.

Kaufman, Leonard, and Peter J Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analy-sis*. Volume 344. John Wiley & Sons.

Loga, Tobias, N Diefenbach, and B Stein. 2012. Typology approach for building stock energy assessment. Main results of the TABULA project.

MacQueen, James, et al. 1967. “Some methods for classification and analysis of multivariate observations.” *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol-ume 1. Oakland, CA, USA., 281–297.

Majcen, D, LCM Itard, and H Visscher. 2013. “Theoretical vs. actual energy consumption of labelled dwellings in the Netherlands: Discrepancies and pol-icy implications.” *Energy policy* 54:125–136.

Mastrucci, Alessio, Olivier Baume, Francesca Stazi, and Ulrich Leopold. 2014. “Estimating energy savings for the residential building stock of an entire city: A GIS-based statistical downscaling approach applied to Rotterdam.” *Energy and Buildings* 75:358–367.

Mata, ´E, A. Sasic Kalagasidis, and F. Johnsson. 2014. “Building-stock aggregation through archetype buildings: France, Germany, Spain and the UK.” *Building and Environment* 81:270–282.

Recast, EPBD. 2010. “Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings (recast).” *Official Journal of the European Union* 18 (06): 2010.

Rousseeuw, Peter J. 1987. “Silhouettes: a graphical aid to the interpretation and validation of cluster analy-sis.” *Journal of computational and applied mathe-matics* 20:53–65.

Sarac¸li, Sinan, Nurhan Do˘gan, and ˙Ismet Do˘gan. 2013. “Comparison of hierarchical cluster analysis meth-ods by cophenetic correlation.” *Journal of Inequali-ties and Applications* 2013 (1): 203.

Schaefer, Aline, and Enedir Ghisi. 2016. “Method for obtaining reference buildings.” *Energy and Build-ings* 128:660–672.

Sokol, Julia, Carlos Cerezo Davila, and Christoph F Reinhart. 2017. “Validation of a Bayesian-based method for defining residential archetypes in urban building energy models.” *Energy and Buildings* 134:11–24.

Sousa Monteiro, Claudia, Carlos Cerezo, Andre´ Pina, and Paulo Ferr˜ao. 2015. “A method for the gener-ation of multi-detail building archetype definitions: Application to the city of Lisbon.” *Proceedings of In-ternational Conference CISBAT 2015 Future Build-ings and Districts Sustainability from Nano to Urban Scale*. LESO-PB, EPFL, 901–906.

The Buildings Performance Institute Europe, (BPIE). 2018. Data Hub for energy performance of build-ings.

Torabi Moghadam, Sara, Guglielmina Mutani, and Patrizia Lombardi. 2016. “GIS-Based Energy Consumption Model at the Urban Scale for the Building Stock.” *EUR*, pp. 56–63.

US Department of Energy, (DOE). 2018. Commercial Reference Buildings.

NOMENCLATURE

k	number of cluster
$MinPts$	minimum points
Eps	epsilon
p	data points
C_i	ith cluster