Energy & Buildings 206 (2020) 109557

Contents lists available at ScienceDirect

Energy & Buildings

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

Office representatives for cost-optimal energy retrofitting analysis: A novel approach using cluster analysis of energy performance certificate databases

Marta Gangolells*, Miquel Casals, Jaume Ferré-Bigorra, Núria Forcada, Marcel Macarulla, Kàtia Gaspar, Blanca Tejedor

Universitat Politècnica de Catalunya, Group of Construction Research and Innovation (GRIC), C/ Colom, 11, Ed. TR5, 08222 Terrassa (Barcelona), Spain

ARTICLE INFO

Article history:
Received 30 May 2019
Revised 12 September 2019
Accepted 26 October 2019
Available online 28 October 2019

Keywords:
Cluster analysis
k-means
Reference buildings
Offices
Energy performance certificates

ABSTRACT

A large number of buildings must be evaluated to formulate energy retrofitting policies for existing building stock. In this context, it is crucial to identify reference buildings that can effectively represent the entire stock, since such buildings can then be used to assess the individualized cost-effectiveness of retrofitting measures. This paper presents a novel approach for identifying and defining a set of reference buildings by applying the *k*-means clustering method to energy performance certificate databases. To this end, a four-step methodology has been envisaged. First, an energy performance certificate database is prepared and variables that have an impact on energy consumption are pre-selected. Selected data are then pre-processed. Next, the *k*-means clustering method is applied. Finally, the resulting cluster centroids are used to identify the closest energy performance certificates in the database, in other words, the representative buildings that will then be used for cost-optimal retrofitting analysis. The methodology is illustrated using the energy performance certificate database managed by the Catalan Institute of Energy (ICAEN), which includes a sample of 13,701 offices. Due to the large number of missing values in the database, the *k*-means clustering algorithm was finally performed over 6,083 energy performance certificates. Seven representative office blocks and offices in industrial buildings and nine representative offices in residential buildings were identified. The results establish the basis for supporting strategic decision-making for energy saving retrofit interventions in existing Spanish offices.

1. Introduction

With the ultimate objective of enhancing the renovation rate of European buildings, both the Energy Performance of Buildings Directive [8] and the Energy Efficiency Directive [9] required Member States to establish a comparative methodological framework for calculating cost-optimal levels of minimum energy performance requirements for buildings and building elements. Since the building sector's heterogeneity hinders the identification of individualized cost-optimal levels, Commission Delegated Regulation (EU) No 244/2012 [11] established the basis for defining reference buildings that can effectively represent the building stock. Reference buildings are useful tools for evaluating energy saving measures in the entire building stock [26] and assessing their energy saving potential, considering a trade-off between the reliability of the results and the required computational effort.

According to the guidelines accompanying Commission Delegated Regulation (EU) No 244/2012 [12], two main approaches have been adopted to identify reference buildings. In the first approach, a real building is selected that represents the most typical building in a specific category. According to Ballarini et al. [2], the real building can be defined using the real example building approach, in which the most representative building is selected by a panel of experts, or the real average building approach, in which statistical building data are analysed to identify a building that has similar characteristics to the mean of the sample. In the second approach, a virtual building is created with the most commonly used materials and systems using statistically significant properties.

The identification of representative buildings using traditional statistical approaches may be biased towards a particular feature (i.e. energy intensity). Consequently, other building characteristics

* Corresponding author.
E-mail address: marta.gangolells@upc.edu (M. Gangolells).

https://doi.org/10.1016/j.enbuild.2019.109557

Nomenclature

B3, C2, C3, D1, D2, D3, E1	climate zone based on winter climate severity (identified by a letter, being A for those locations with the warmest winter and E for those locations with the coldest winter) and summer climate severity (identified by a number, being 1 for those locations with the mildest summer and 4 for those locations with the hottest summer)
CALENER GT	general procedure for the energy rating of large tertiary sector buildings recognized by the Spanish government
CALENER VYP	general procedure for the energy rating of dwellings and small tertiary sector buildings recognized by the Spanish government
CE3	simplified procedure for the energy rating of existing buildings recognized by the Spanish government developed by Applus Norcontrol S.L.U.
CE3X	simplified procedure for the energy rating of existing buildings recognized by the Spanish government developed by Natural Climate Systems S.A.
CERMA	simplified procedure for the energy rating of residential buildings recognized by the Spanish government
DHW	domestic hot water
EOFF	individualized sustainable and cost optimal energy retrofitting solutions for the existing office building stock
HULC	general procedure for the energy rating of dwellings and tertiary sector buildings recognized by the Spanish government (Lider Calener unified tool)
ICAEN	Catalan Institute of Energy
NBE-CT79	compulsory basic building norm regarding thermal conditions in buildings
OB1-OB7	cluster codes for data subset 1 including office blocks and offices in industrial buildings
OR1-OR9	cluster codes for data subset 2 including offices in residential buildings
RMSSTD	root-mean-square standard deviation
SCS	summer climate severity
TBC	Technical Building Code
WCS	winter climate severity

that are important for building energy modelling (i.e. geometry, occupancy, technologies and total energy consumption) may be underrepresented [28]. In contrast, when traditional statistical approaches are used based on data filtering of highly correlated variables, datasets typically result in a high number of segments. These shortcomings can be easily overcome using clustering techniques. Clustering is a data-mining approach that identifies homogeneous groups of objects in a dataset [19]. The most commonly used algorithm for building energy analysis is *k*-means clustering [3]. *k*-means is a partitional clustering technique that assigns objects to clusters to minimize the distance from objects to the cluster centre, once a user-defined number of clusters has been selected [23].

The choice of an existing representative building always requires a large amount of information on the building stock [4]. This is especially true when clustering techniques are used as they require complete datasets [28]. Previous research initiatives in the field of reference building identification applied clustering methods using databases to correlate building, construction and operational characteristics with energy consumption data (Table 1). However, databases were mostly created using processes that require high time and resource consumption, such as surveys or energy audits. As a result, the databases contained a limited number of samples and features. Since the implementation of the recast Energy Performance Building Directive [8], energy performance certificates have provided an excellent source of information on the characteristics of the building stock [5] and they have great potential for applications based on data science [24]. This approach is further legitimized by the latest revision of the Energy Performance Building Directive [10] that includes provisions to ensure high-quality data on building stock through energy performance certificate databases.

Approaches used to date mostly focus on schools, residential buildings and mixed-use buildings (Table 1). Offices have not been investigated in many studies, even though office buildings have much higher energy consumption per capita than residential buildings because they are equipped with energy-intensive equipment and have high space cooling and heating demands [7].

The main objective of this paper is to develop a methodology for identifying a limited set of real reference buildings that are representative of the entire building stock. The novelty of this contribution lies in applying the *k*-means clustering algorithm over a large dataset including information retrieved from energy performance certificates. The case study illustrating the methodology focuses on the subset of Spanish offices, representing a progress beyond the state of the art because existing approaches mostly focus on other building typologies with lower energy consumption. Analysis of representative offices will provide office managers, construction practitioners and other relevant stakeholders with a better understanding of how to cost-effectively promote energy conservation in current office stock. In addition, the development of retrofit strategies for offices that react similarly to energy efficiency measures may become essential to support the development of future policies and funding schemes. Section 2 presents the methodology used in this research, whereas Section 3 reports the results. Finally, conclusions and future work are presented in Section 4.

2. Methodology

The methodology developed in this paper (Fig. 1) has four steps:

Step 1: Database preparation and variables preselection

This step involves gathering, combining, integrating, structuring and organizing the energy performance certificate database to ensure it is ready for the subsequent analysis. Taking into account that the implementation of energy performance certification schemes varies a lot across countries and even regions, preparation steps largely depend on the initial condition of the energy

Table 1
Review of the approach used for representative building identification using clustering techniques.

Reference	Typology of buildings	Number of database entries
Tardioli et al. [28]	Mixed-use buildings	9,500
Deb and Lee [7]	Office buildings	56
Ghiassi and Mahdavi [19]	Mixed-use buildings	750
Schaefer et al. [26]	Residential	103
Pieri et al. [25]	Hotels	35
Arambula Lara et al. [1]	Schools	60
Famuyibo et al. [13]	Residential	150
Gaitani et al. [14]	Schools	1,100

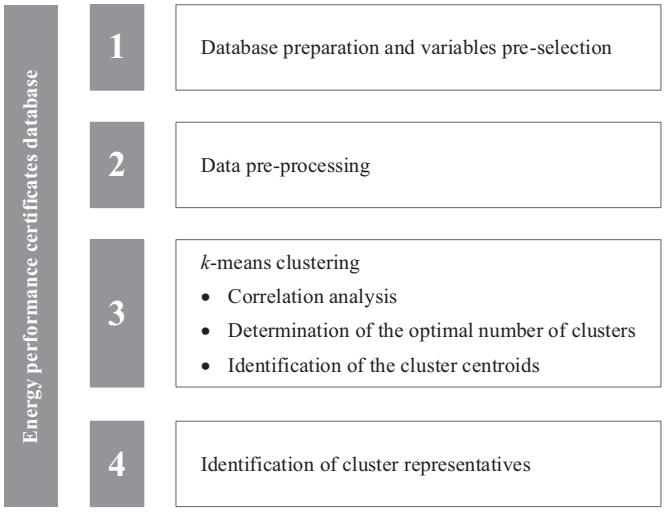

Fig. 1. Methodology developed in this research.

performance certificates databases. Database variables that can have an impact on energy consumption are then pre-selected.

Step 2: Data pre-processing

Concerns about data quality in energy performance certificate databases [5,24] make data pre-processing crucial to ensure the robustness of the analysis. As proposed by Pasichnyi et al. [24], it is vital to perform consistency checks including constraint rules for specific columns and physical rules involving the analysis of values in several columns. Data enrichment may also be performed if needed by creating additional variables using existing information in the database. Categorical variables must be converted to dummy variables. Normalization is also necessary because energy performance certificate databases usually include variables whose magnitudes differ greatly. Variables can be rescaled (0–1) using min-max normalization.

Step 3: *k*-means clustering

The *k*-means clustering technique divides n observations (in this case, energy performance certificates) into k non-overlapping clusters (representative buildings) so that each observation belongs to the cluster with the closest centroid. Along the lines of Casals et al. [6] and as expressed in Eq. (1), k centres c are chosen to minimise ϕ:

$$\phi = \sum_{x \in X} \min_{c \in C} \| x - c \|^2 \quad (1)$$

Where:

ϕ is the total squared distance between each point and its closest centre, x represents a data point, X is the set of data points of the pre-processed energy performance certificate database, c denotes the cluster centre and C is the set of cluster centres (the clustering).

Step 3.1: Correlation analysis

To effectively characterize the sample, original data needs to be weighted according to their contribution to the buildings' energy performance. Correlation between the annual non-renewable primary energy consumption per square metre (including heating, cooling, domestic hot water and lighting) and the pre-selected variables affecting the energy consumption of buildings is investigated through forward stepwise regression analysis, as suggested by Gao and Malkawi [17]. In the forward approach, the model starts with no variables in it. During the first step, the most significant variable is added to the model. At each subsequent step, the most significant variable of all remaining variables is selected and introduced to the model. This process is repeated until there are no more variables that meet the criterion set by the user, in this case, a maximum p-value of 0.05. Afterwards and within the backwards approach, the variable with the highest p-value is removed at each step. This is repeated until no further variables can be deleted because their p-value is below the pre-defined exit tolerance (0.10 in this case).

Correlation analysis requires the exclusion of all energy performance certificates that have a missing value in at least one of the pre-selected variables. Considering that missing values can be spread through the complete dataset, the final database may include few complete entries. If there are many missing values in the database, regression coefficients to weigh variables in the *k*-means clustering can be obtained using a two-step approach. The analysis is first performed considering all pre-selected variables, and later, considering only variables that are found to be significant in the first analysis. Thus, the analysis is performed over a higher number of rows without any missing values.

Step 3.2: Determination of the optimal number of clusters

As recognized by the guidelines accompanying Commission Delegated Regulation (EU) No 244/2012 [11], building stock is reflected more realistically with a higher number of reference buildings, but there is obviously a trade-off between the representativeness of the building stock and the resources available for the performance assessment. As stated by Halkidi et al. [20] and for non-hierarchical clustering, the optimum number of clusters is determined using the root-mean-square standard deviation (RMSSTD) validity index (Eq. (2)).

$$RMSSTD = \left[\frac{\sum_{\substack{i=1\ldots k \\ j=1\ldots d}} \sum_{q=1}^{n_{ij}} \left(x_q - \overline{x_j}\right)^2}{\sum_{\substack{i=1\ldots k \\ j=1\ldots d}} \left(n_{ij} - 1\right)} \right]^{\frac{1}{2}} \quad (2)$$

Where k is the number of clusters, d represents the number of variables or the data dimension, n_{ij} corresponds to the number of

data values of j dimension that belong to cluster i and $\overline{x_j}$ is the mean of data values of j dimension.

The RMSSTD decreases as the number of clusters (k) increases. In fact, RMSSTD is 0 when there are as many clusters (k) as data points in the dataset, because then each data point is the centroid of its own cluster and there is no error between them. The goal is to identify a small number of clusters (k) that still provide a small RMSSTD. Therefore, according to the Elbow method, the RMSSTD is computed and plotted for a range of k values. The inflection point in the plot (a graph angle also known as the elbow) indicates that adding more clusters does not significantly improve the data modelling.

Step 3.3: Identification of cluster centroids

Cluster centroids are obtained by selecting the mean value for each feature in each cluster.

Step 4: Identification of cluster representatives

To avoid using a virtual representative building for each cluster, the Euclidean distance between the centroid of the cluster and all the energy performance certificates in the cluster is calculated. The closest building, that is, the energy performance certificate with the smallest distance to the cluster centroid, is selected as the cluster representative (Eq. (3)).

$$b_{ref,k} = argmin_q ||b_q - c_k||^2 \quad (3)$$

Where, $b_{ref,k}$ is the representative energy performance certificate of the k cluster, b_q represents the energy performance certificate q and c_k denotes the centre of the cluster k.

The representativeness index (Eq. (4)) can be used to estimate the level of similarity between the cluster centroid and the selected energy performance certificate.

$$Representativeness\ (\%) = \frac{\sum_{i=1}^{n} k_i \cdot \left(1 - |p_{i,centroid} - p_{i,certificate}|\right)}{\sum_{i=1}^{n} k_i} \cdot 100 \quad (4)$$

Where, k denotes the weighting coefficient (extracted from the regression analysis) for a given parameter i, $p_{i,centroid}$ represents the value of the parameter i in the cluster centroid and $p_{i,certificate}$ is the value of the parameter i in the selected energy performance certificate.

One hundred per cent representativeness indicates that the selected energy performance certificate has the same characteristics in highly correlated variables as those of the cluster centroid.

3. Results

The following subsections describe the application of the methodology to the energy performance certificate database, including 718,872 energy performance certificates, 13,701 of which are related to offices, collected in Catalonia (north-east of Spain) by the Catalan Institute of Energy (ICAEN) between the entry into force of Royal Decree 235/2013 [27] in June 2013 and July 2018. According to the General Directorate for the Land Registry [18], Catalonia currently has 47,212 offices. So as to check the representativeness of the sample, the minimum sample size is calculated according to Eq. (5), developed by Krejcie and Morgan [21].

$$Minimum\ required\ sample\ size = \frac{\chi^2 \cdot N \cdot P(1-P)}{d^2(N-1)} + \chi^2 \cdot P(1-P) \quad (5)$$

Where: χ^2 is the table value of chi-square for one degree of freedom at the desired confidence level ($\chi^2 = 3.841$ considering a confidence level of 95%), P is the population proportion (most adverse case of $P = 0.50$ is considered), N is the population size and d is the degree of accuracy expressed as a proportion ($d = 0.05$ considering an accuracy of 5%).

As the minimum sample size for a population of 47,212 is 381 offices, the sample (in other words, the number of entries in the energy performance certificate database) is deemed appropriate and representative.

3.1. Database preparation and variables pre-selection

The ICAEN database was comprised of three anonymized dump files with several tables. In most of the tables, each row corresponds to a particular energy performance certificate and each column stands for a particular variable in the energy performance certificate dataset. Original database files were gathered and organized into a single complete table, from which duplicated and non-relevant features were eliminated. The resulting dataset was characterized by over 150 features including general administrative data (i.e. reference id, address, climate zone, etc.), information related to energy performance (label, energy certification procedure and detailed information on energy demand, energy consumption and emissions), main characteristics of the office including geometry (i.e. useful floor area), thermal envelope (including both the opaque closures and openings) and existing facilities (i.e. heating, cooling and domestic hot water). Oracle dump files were later exported into IBM SPSS Statistics v25.0 and KNIME Analytics Platform 3.7.0 for analysis.

Regarding the energy certification procedure, Royal Decree 235/2013 [27] allows using either a general or a simplified procedure. General procedures require developing a graphical model of the building and therefore, a higher amount of building data is needed. General procedures usually provide more accurate results but require higher skills. HULC (replacing the old Calener VYP and Calener GT since 2016) is the reference software for energy certification using the general procedure in Spain. Other general procedures recently endorsed by the Spanish Ministry for the Ecological Transition are CYPE-THERM and SG Save (2018). Simplified procedures are a reduction of the reference method requiring less data. In this case, inputs are introduced using data entry tabs. Simplified procedures currently validated by the Spanish government are CE3X, CE3 and CERMA.

A preliminary analysis of the database revealed that most of the energy performance certificates (76.58%) were related to offices located either at street level or in higher floors of residential buildings, whereas the remaining 23.42% were found to be related to office blocks and offices in industrial buildings. Offices located in residential buildings tend to be small (between 40.60 m^2 and 216.06 m^2) and the corresponding energy performance certificates are mostly obtained using simplified procedures (99.83%). Typical examples include doctors' offices, dental clinics, lawyers' offices, real estate agencies, agencies undertaking administrative work, small consultancy firms, etc. The energy performance of this kind of offices and the impact that energy retrofitting actions may have is not expected to be much different from that of individual dwellings in multi-family blocks. In contrast, office blocks and offices in industrial buildings are larger (between 48.00 m^2 and 1,039.10 m^2). In this case, energy performance certificates are also mostly obtained using simplified procedures (93.48%) but the proportion of energy performance certificates obtained using general procedures is higher (6.52%). In addition, energy retrofitting action may be of greater magnitude and different from what would be expected for offices located in residential buildings. Consequently, the dataset entries were filtered using information related to office configuration, leading to data subset 1 (office blocks and offices in

Table 2
Pre-selected variables.

Variables	Adopted values
Energy performance certification procedure	Simplified procedures such as CE3X, CE3 and CERMA or general procedures such as HULC, CALENER VYP and CALENER GT [22]
Annual non-renewable primary energy consumption per square metre	Numerical value expressed in kWh_p/m^2·year
Useful floor area	Numerical value expressed in m^2
Shape factor	Relation between the building thermal envelope and the building volume, expressed in m^2/m^3
Solar contribution for domestic hot water	Numerical value expressed as a percentage of the DHW (domestic hot water) demand heated with solar thermal collectors
Existence of photovoltaic energy	Yes or no
Existence of geothermal energy	Yes or no
Building norms	None for buildings built before 1980, NBE-CT 79 for buildings built from 1980–2006 and TBC (Technical Building Code) for buildings built after 2006 [15]
Climate zone	Based on WCS (winter climate severity), identified by a letter, and SCS (summer climate severity), identified by a number: B3, C2, C3, D1, D2, D3 or E1 [16]
Office type	Office block, office in an industrial building or office in a residential building
Heating system	No heating, individual or centralized
Cooling system	No cooling, individual or centralized
Existence of thermal insulation in building envelopes	Yes, no or unknown
Window glazing type	Single glazing, double glazing or low emissivity double glazing
Heating energy source	Natural gas, propane, liquefied petroleum gas, biomass, electricity, diesel oil, carbon or no heating
Cooling energy source	Natural gas, electricity, diesel oil, carbon or no cooling
Domestic hot water energy source	Solar, propane, butane, diesel oil, natural gas, liquefied petroleum gas, electricity or biomass

Table 3
Variables and threshold values used for detecting errors in the energy performance certificates database.

Variable	Value threshold
Useful floor area (S_u)	$S_u \geq 10$ m^2
Clear height (h)	$2.2\ m \leq h \leq 5.0$ m
Shape factor (SF)	$0.01\ m^2/m^3 \leq SF \leq 2.30\ m^2/m^3$
Thermal enclosure area (S_{te})	$S_{te} \geq 5$ m^2
Non-renewable primary energy consumption (E_p)	$24.4\ kWh_p/m^2$·year $\leq E_p \leq 1{,}000\ kWh_p/m^2$·year

Table 4
Conversion of the categorical variable building norm into its equivalent dummy variables.

Categorical variable		Dummy variables	
		NBE-CT 79	TBC
Building norm	None	0	0
	NBE-CT 79	1	0
	TBC	0	1

industrial buildings) and data subset 2 (offices in residential buildings).

Table 2 summarizes the database variables that were considered to have an impact on energy consumption.

3.2. Data pre-processing

Data pre-processing included cleaning values outside the allowed ranges according to constraint rules and physical limitations (Table 3). Data entry errors were also detected, and if possible corrected or otherwise deleted. Some textual values had to be standardized because the same concept can be described using different words when data are entered in energy performance certificate forms.

Nominal variables were decomposed and converted to dummy variables. Table 4 exemplifies the conversion of a categorical variable composed of three categories into two dummy variables. Finally, both continuous and interval variables were normalized to the same magnitude (0–1), using a min-max normalization, to avoid deviations caused by variables with large variation ranges that could lead to misleading cluster results. After data pre-processing, the original database was reduced to 13,076 energy performance certificates for offices.

3.3. k-means clustering

The *k*-means clustering technique was applied to the two data subsets: (i) data subset 1 including office blocks and offices in industrial buildings and (ii) data subset 2 including offices in residential buildings. Data were clustered using the KNIME Analytics Platform 3.7.0.

3.3.1. Correlation analysis

To weight original data, a correlation analysis between non-renewable primary energy consumption (expressed in kWh_p/m^2·year) and the pre-selected variables was performed with IBM SPSS Statistics v25.0. Due to the large number of missing values in the database and along the lines of what was suggested in step 3.1 of the methodology, the two-step approach was used. For data subset 1, when all the pre-selected variables were considered, only 843 energy performance certificates could be used. The remaining energy performance certificates were rejected because they had at least one missing value in one of the pre-selected variables. During the second step, only variables that were found to be significant in the first correlation analysis were taken into account, which increased the number of rows without missing values to 2,501. For data subset 2, the two-step approach increased the number of analysed energy performance certificates from 1,984 to 3,582. Regression coefficients obtained in the second iteration were used to weigh the variables in the *k*-means clustering (Tables 5 and 6).

Table 5
Regression coefficients obtained in the second iteration for normalized data subset 1.

Predictor	Regression coefficient
Building norm	0.319
Window glazing type	0.238
Climate zone	0.217
Cooling system	0.098
Useful floor area	0.090
Shape factor	0.038

Table 6
Regression coefficients obtained in the second iteration for normalized data subset 2.

Predictor	Regression coefficient
Shape factor	0.407
Domestic hot water energy source	0.282
Climate zone	0.121
Building norm	0.085
Cooling system	0.061
Heating energy source	0.025
Existence of thermal insulation in building envelopes	0.020

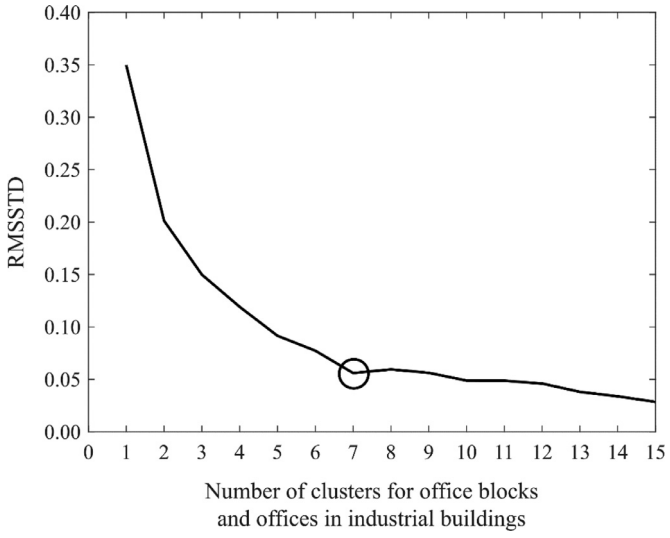

Fig. 2. Optimal number of clusters for data subset 1.

3.3.2. Determination of the optimal number of clusters

According to the Elbow method, *k*-means clustering was run on the datasets for a range of values of *k* (from 1 to 15) and the corresponding root-mean-square standard deviation (RMSSTD) was computed and plotted (Figs. 2 and 3). Considering that lines in the charts look like an arm, the elbows are marked with a small circle. The appropriate number of clusters was found to be seven for the data subset including block offices and offices in industrial buildings (Fig. 2) and nine for the data subset of offices in residential buildings (Fig. 3).

3.3.3. Identification of cluster centroids

The 2,501 energy performance certificates for office blocks and offices in industrial buildings in the database were grouped into 7 clusters (OB1-OB7), with different benchmarking references (in terms of annual non-renewable primary energy consumption per square metre) and representativeness (Table 7). Cluster OB3 represents a higher number of offices (29.51%), followed by clusters OB1 and OB6, with representativeness of 22.35% and 15.31% respectively.

The 3,582 energy performance certificates for offices in residential buildings were clustered into nine groups (OR1-OR9). In this

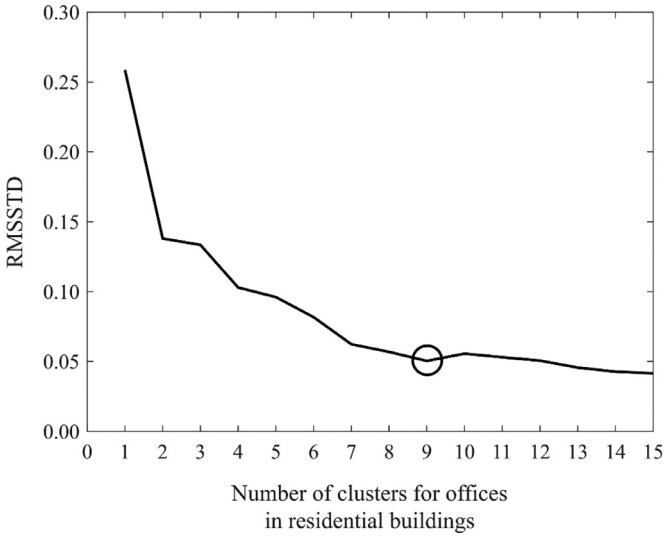

Fig. 3. Optimal number of clusters for data subset 2.

Table 7
Benchmarking reference and representativeness of the identified clusters.

Cluster code	Centroid (kWh_p/m^2·year)	Number of buildings	
		(ut.)	(%)
Office blocks and offices in industrial buildings			
OB1	239.79	559	22.35
OB2	248.25	352	14.07
OB3	266.83	738	29.51
OB4	218.63	126	5.04
OB5	282.01	98	3.92
OB6	156.90	383	15.31
OB7	288.28	245	9.80
Offices in residential buildings			
OR1	270.37	247	6.90
OR2	269.97	30	0.84
OR3	253.30	1,520	42.43
OR4	260.55	389	10.86
OR5	257.88	576	16.08
OR6	167.43	137	3.82
OR7	238.72	193	5.39
OR8	231.54	345	9.63
OR9	219.50	145	4.05

case, and as shown in Table 7, cluster OR3 represents a higher number of offices in the database (42.43%), followed by cluster OR5 (16.08%) and cluster OR4 (10.86%).

The main characteristics of the centroids are summarized in Table 8 (office blocks and offices in industrial buildings) and in Table 9 (offices in residential buildings), using the variables that were found to be significant in each case (Tables 5 and 6, respectively). The representative office at the centroid of a cluster is not an existing office but serves as the reference for all the offices in the same cluster.

3.4. Identification of cluster representatives

The following tables summarize the representativeness of the best cluster representatives for office blocks and offices in industrial buildings (Table 10) and for offices in residential buildings (Table 11). Considering that clusters are defined to subsequently assess the cost-effectiveness of energy retrofitting actions in office stock, energy performance certificates performed with the general procedure (HULC, CALENER GT or CALENER VYP) are always

Table 8
Main characteristics of each of the identified cluster centroids for data subset 1.

Cluster code	Building norm	Window glazing type	Climate zone	Cooling system	Useful floor area (m^2)	Shape factor (m^2/m^3)
OB1	None	Simple glazing	C2	Individual	312.71	0.33
OB2	None	Double glazing	C2	Individual	507.98	0.31
OB3	NBE-CT 79	Double glazing	C2	Individual	518.47	0.36
OB4	NBE-CT 79	Simple glazing	C2	No cooling	114.75	0.38
OB5	NBE-CT 79	Simple glazing	D2	Individual	178.36	0.43
OB6	TBC	Low emissivity double glazing	C2	Individual	1,323.39	0.36
OB7	NBE-CT 79	Simple glazing	C2	Individual	196.01	0.39

Table 9
Main characteristics of each of the identified cluster centroids for data subset 2.

Cluster code	Shape factor (m^2/m^3)	Domestic hot water energy source	Climate zone	Building norm	Cooling system	Heating energy source	Existence of thermal insulation in building envelopes
OR1	0.34	Electricity	D2	None	Individual	Electricity	No
OR2	0.32	Butane	C2	None	No	No heating	No
OR3	0.35	Electricity	C2	None	Individual	Electricity	No
OR4	0.34	Electricity	D2	NBE-CT 79	Individual	Electricity	Yes
OR5	0.49	Electricity	C2	NBE-CT 79	Individual	Electricity	Yes
OR6	0.08	Electricity	C2	None	Individual	Electricity	No
OR7	0.16	Electricity	C2	NBE-CT 79	Individual	Electricity	Yes
OR8	0.27	Natural gas	C2	None	Individual	Electricity	No
OR9	0.35	Electricity	C2	TBC	Individual	Electricity	Yes

Table 10
Energy performance certification procedure and representativeness of the selected energy performance certificate for the centroids of data subset 1.

Cluster code	Energy performance certificate			Representativeness (%)
	Id	Procedure		
OB1	WSH0RLX5J	General	CALENER VYP	89.83
	492YV4MQQ	Simplified	CE3X	99.90
OB2	ZT8GRBMTZ	General	CALENER GT	86.83
	NBMVDF9ZH	Simplified	CE3X	99.97
OB3	HJGNBC8LM	General	CALENER GT	89.49
	N0T8YCJH	Simplified	CE3X	99.94
OB4	–	General	–	–
	144GP737Z	Simplified	CE3X	99.90
OB5	1MP0SZRTM	General	HULC (CALENER VYP)	76.07
	763YX3PDJ	Simplified	CE3X	99.94
OB6	GXNGZDVK6	General	CALENER VYP	99.70
	6Q8QFD0NB	Simplified	CE3X	99.38
OB7	–	General	–	–
	W5PG2NG6J	Simplified	CE3X	99.94

Table 11
Energy performance certification procedure and representativeness of the selected energy performance certificates for the centroids of data subset 2.

Cluster code	Energy performance certificate			Representativeness (%)
	Id	Procedure		
OR1	JPT21FZGY	Simplified	CE3X	99.97
OR2	R9F2QBJTW	Simplified	CE3X	98.75
OR3	79CS4XP1D	Simplified	CE3X	99.99
OR4	MD50Y1ZLP	Simplified	CE3X	99.38
OR5	2H1PLZG4D	Simplified	CE3X	99.98
OR6	56Q5CPS34	Simplified	CE3X	99.20
OR7	9T2DBSP8L	Simplified	CE3X	99.96
OR8	16JSZTR56	Simplified	CE3X	99.91
OR9	6Q2JTGQ2J	Simplified	CE3X	99.98

preferred over those using simplified procedures (CE3X, CE3 and CERMA) as they provide more accurate results in building energy performance simulation [22]. Table 10 shows, for each cluster of data subset 1, the representativeness of the closest energy performance certificate performed with the general procedure. However, in some cases (OB4 and OB7), the clusters do not include any energy performance certificates obtained with the general procedure (Table 10). In these cases, energy performance certificates obtained with the simplified procedure will need to be used. However, it must be noted that energy performance certificates obtained using simplified procedures generally have higher levels of representativeness in relation to the theoretical cluster centroid (Table 10). Table 11 shows the selected representatives of data subset 2, taking into account that in this case almost all the energy performance certificates in the database were performed using simplified procedures.

To better illustrate the research findings summarized in Tables 7, 8, 9, 10 and 11, results related to cluster OB1 are fully described here. A total of 559 office blocks and offices in an industrial building (22.35%) were found to fall within cluster OB1 (Table 7). The representative office at the centroid of cluster OB1 has non-renewable primary energy consumption of 239.79 kWh_p/m^2·year (Table 7). As shown in Table 8, this theoretical office is in climate zone C2 and was built before 1979 without meeting any thermal regulations. With a useful floor area of 312.71 m^2 and a shape factor of 0.33 m^2/m^3, windows are simple-glazed and there is an

Table 12
Main characteristics of office representatives for cluster OB1.

	Cluster centroid	Energy performance certificate	
		Certificate id WSH0RLX5J	Certificate id 492YV4MQQ
Certification procedure	–	CALENER VYP	CE3X
Non-renewable primary energy consumption (kWh_p/m^2·year)	239.79	69.84	238.53
Building norm	None	None	None
Window glazing type	Simple glazing	Simple glazing	Simple glazing
Climate zone	C2	C2	C2
Cooling system	Individual	No cooling	Individual
Useful floor area (m^2)	312.71	193.51	116.87
Shape factor (m^2/m^3)	0.33	0.93	0.34

individual cooling system (Table 8). Among all the energy performance certificates in the database performed with the general procedure, the closest real office to the OB1 centroid was found to be that with the id WSH0RLX5J (Table 10). This energy performance certificate was obtained using CALENER VYP and shows 89.83% similarity to the representative office at the centroid of cluster OB1 (Table 10). Among all the energy performance certificates obtained using a simplified procedure, the closest was found to be id 492YV4MQQ, with 99.90% representativeness in relation to the theoretical representative office for the OB1 cluster (Table 10). As indicated in Table 10, this energy performance certificate was obtained using CE3X software. Table 12 summarizes the main characteristics of the best-fitting energy performance certificates for cluster centroid OB1. This energy performance certificates could now be used for cost-optimal energy retrofitting analysis and the results would be representative of all the office blocks in cluster OB1.

Adoption of the clustering approach for the energy performance certificates dataset resulted in a total of 16 office representatives. However, a traditional statistical approach based on the highly correlated variables would have resulted in a matrix with a total of 3,159 segments. In this case, office blocks and offices in industrial buildings would have been segmented into 972 categories, considering the product of six significant variables including building norms, window glazing type, climate zone, cooling system, useful floor area and shape factor. Offices in residential buildings would have been represented by 2,187 segments, considering a matrix with 7 significant variables including shape factor, DHW (domestic hot water) energy source, climate zone, building norms, cooling system, heating energy source and existence of thermal insulation in building envelopes. Although some techniques such as frequency histograms or use of representative parameters of building regulations could help to reduce the number of possible combinations [13], such a large number of reference offices would obviously hinder office stock description and the subsequent analysis of energy retrofitting measures.

4. Conclusions

This paper presented replicable methodology to identify a limited set of buildings that are representative of the entire stock by clustering the information in the energy performance certificates databases. The methodology can be applied to national databases for any particular building typology and corresponding reference buildings can be obtained. The approach is especially useful within the framework of the Energy Performance of Buildings Directive [8], since the impact of retrofit measures can be modelled using a cost-optimal approach for all the buildings in a cluster with a limited computational effort. The methodology can also contribute to examining buildings' energy baseline and to better estimate the energy saving potential of retrofitting actions in the stock.

To demonstrate the applicability of the methodology, the clustering based grouping methodology was applied to identify a limited set of office representatives using a large energy performance certificate database containing 13,701 entries and over 150 features. After excluding all energy performance certificates that had a missing value in at least one of the seventeen pre-selected variables, the *k*-means clustering was applied to 6,083 energy performance certificates. The results identified seven cluster representatives of offices blocks and offices in industrial buildings and nine cluster representatives of offices in residential buildings. Representativeness was found to range between 76.07% and 99.99%, depending on the energy performance certification procedure.

Clustering the energy performance certificate databases has proved to be a useful way to obtain a limited number of reference buildings that are representative of the entire dataset. Clustering guarantees that groups are created considering not only the energy use intensity index but other building related variables that have an impact on energy consumption. The limitations of this approach are mainly derived from aspects related to data availability and accuracy in the energy performance certificates database. Smart metering roll-out is expected to provide large quantities of real data that, if linked to the energy performance certificate database, show great potential for improving building stock modelling.

Declaration of Competing Interest

None.

Acknowledgments

This research was supported by the Spanish Ministry of Economy, Industry and Competitiveness under R&D project EOFF, reference no. BIA2016-75382-R (AEI/FEDER, UE). The Catalan Institute of Energy (ICAEN) is gratefully acknowledged for allowing access to the energy performance certificate database.

References

[1] R. Arambula Lara, G. Pernigotto, F. Cappelletti, P. Romagnoni, A. Gasparella, Energy audit of schools by means of cluster analysis, Energy Build. 95 (2015) 160–171.
[2] I. Ballarini, S.P. Corgnati, V. Corrado, Use of reference buildings to assess the energy saving potentials of the residential building stock: the experience of Tabula project, Energy Policy 68 (2014) 273–284.
[3] E.H. Borgstein, R. Lamberts, J.L.M. Hensen, Evaluating energy performance in non-domestic buildings: a review, Energy Build. 128 (2016) 734–755.
[4] A. Brandão de Vasconcelos, M.D. Pinheiro, A. Manso, A. Cabaço, A Portuguese approach to define reference buildings for cost-optimal methodologies, Appl. Energy 140 (2015) 316–328.
[5] Buildings Performance Institute Europe, Energy performance certificates across the EU, A Mapp. Nat. Approaches (2014) Available at http://bpie.eu/uploads/lib/document/attachment/81/BPIE_Energy_Performance_Certificates_EU_mapping_-_2014.pdf . Accessed on 08 March 2019.
[6] M. Casals, M. Gangolells, N. Forcada, M. Macarulla, Reducing lighting electricity use in underground metro stations, Energy Convers. Manag. 119 (2016) 130–141.
[7] C. Deb, S.E. Lee, Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data, Energy Build. 159 (2018) 228–245.

[8] Europe, 2010. Directive 2010/31/EU of the European parliament and of the council of 19 May 2010 on the energy performance of buildings. Available at: <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=celex%3A32010L0031>. Accessed on 05 March 2019.
[9] Europe, 2012. Directive 2012/27/EU of the European parliament and of the council of 25 October 2012 on energy efficiency, amending directives 2009/125/EC and 2010/30/EU and repealing directives 2004/8/EC and 2006/32/EC. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1399375464230&uri=CELEX:32012L0027>. Accessed on 18 March 2019.
[10] Europe, 2018. Directive (EU) 2018/844 of the European parliament and of the council of 30 May 2018 amending directive 2010/31/EU on the energy performance of buildings and directive 2012/27/EU on energy efficiency. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018L0844&from=EN>. Accessed on 05 March 2019.
[11] European Commission, 2012a. Commission Delegated Regulation (EU) no 244/2012 of 16 January 2012 supplementing directive 2010/31/EU of the European parliament and of the council on the energy performance of buildings by establishing a comparative methodology framework for calculating cost-optimal levels of minimum energy performance requirements for buildings and building elements. Available at: <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:081:0018:0036:EN:PDF>. Accessed on 08 March 2019.
[12] European Commission, 2012b. Guidelines accompanying commission Delegated Regulation (EU) no 244/2012 of 16 January 2012 supplementing directive 2010/31/EU of the european parliament and of the council on the energy performance of buildings by establishing a comparative methodology framework for calculating cost-optimal levels of minimum energy performance requirements for buildings and building elements. Available at: <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2012:115:0001:0028:EN:PDF⟩. Accessed on 08 March 2019.
[13] A.A. Famuyibo, A. Duffy, P. Strachan, Developing archetypes for domestic dwellings: an Irish case study, Energy Build. 50 (2012) 150–157.
[14] N. Gaitani, C. Lehmann, M. Santamouris, G. Mihalakakou, Using principal component and cluster analysis in the heating evaluation of the school building sector, Appl. Energy 87 (6) (2010) 2079–2086.
[15] M. Gangolells, M. Casals, Resilience to increasing temperatures: residential building stock adaptation through codes and standards, Build. Res. Inf. 40 (6) (2012) 645–664.
[16] M. Gangolells, M. Casals, N. Forcada, M. Macarulla, E. Cuerva, Energy mapping of existing building stock in Spain, J. Clean. Prod. 112 (2016) 3895–3904.
[17] X. Gao, A. Malkawi, A new methodology for building energy performance benchmarking: an approach based on intelligent clustering algorithm, Energy Build. 84 (2014) 607–616.
[18] General Directorate for the Land Registry, Ministry of finance, Spanish government, Cadastr. Stat. (2018) Available at http://www.catastro.minhap.gob.es/esp/estadisticas.asp . Accessed on 5 December 2018.
[19] N. Ghiassi, A. Mahdavi, Reductive bottom-up urban energy computing supported by multivariate cluster analysis, Energy Build. 144 (2017) 372–386.
[20] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2) (2001) 107–145.
[21] R. Krejcie, D. Morgan, Determining sample size for research activities, Educ. Psychol. Meas. 30 (3) (1970) 607–610.
[22] L.M. López-González, L.M. López-Ochoa, J. Las-Heras-Casas, C. García-Lozano, Energy performance certificates as tools for energy planning in the residential sector. The case of La Rioja (Spain), J. Clean. Prod. 137 (2016) 1280–1292.
[23] X. Li, R. Yao, M. Liu, V. Costanzo, W. Yu, W. Wang, A. Short, B. Li, Developing urban residential reference buildings using clustering analysis of satellite images, Energy Build. 169 (2018) 417–429.
[24] O. Pasichnyi, J. Wallin, F. Levihn, H. Shahrokni, O. Kordas, Energy performance certificates - New opportunities for data-enabled urban energy policy instruments? Energy Policy 127 (2019) 486–499.
[25] S.P. Pieri, I. Tzouvadakis, M. Santamouris, Identifying energy consumption patterns in the Attica hotel sector using cluster analysis techniques with the aim of reducing hotels' CO_2 footprint, Energy Build. 94 (2015) 252–262.
[26] A. Schaefer, E. Ghisi, Method for obtaining reference buildings, Energy Build. 128 (2016) 660–672.
[27] Spain, 2013. Royal Decree 235/2013 approving the basic procedure for certification of energy efficiency of buildings (Real Decreto 235/2013, de 5 de abril, por el que se aprueba el procedimiento básico para la certificación de la eficiencia energética de los edificios). Available at: <http://www.boe.es/boe/dias/2013/04/13/pdfs/BOE-A-2013-3904.pdf>. Accessed on 5 December 2018.
[28] G. Tardioli, R. Kerrigan, M. Oates, J. O'Donnell, D.P. Finn, Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach, Build. Environ. 140 (2018) 90–106.