Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 157 (2019) 1164–1181

Energy Procedia

www.elsevier.com/locate/procedia

Technologies and Materials for Renewable Energy, Environment and Sustainability, TMREES18, 19–21 September 2018, Athens, Greece

True power consumption labeling and mapping of the health system of the Castilla y León region in Spain by clustering techniques

Álvaro de la Puente-Gil[a], Alberto González-Martínez[a], David Borge-Diez[a], Miguel-Ángel Martínez-Cabero[b] and Miguel de Simón-Martín[a,*]

[a]*Energy Resources' Smart Management (ERESMA) Research Group. Dept. Area of Electrical Engineering. Universidad de León. Campus de Vegazana s/n, León, 24071, Spain.*

[b]*Ente Regional de la Energía (EREN) de Castilla y León. Junta de Castilla y León. Avda. de los Reyes Leoneses, 11, León, 24008, Spain.*

Abstract

The latest revisions in April 2018 of the 2010/31/UE and 2012/27/UE Directives on Energy Efficiency and Energy Savings respectively, point out the need of the development of smart energy indexes for buildings with the aim to (i) supervise the energy consumption on the building sector -that currently represents up to one third of the total final energy consumption- and (ii) lead the appropriate actions to transform the current buildings stock to nearly Zero Energy Buildings and Positive Energy Buildings. From public managed buildings, the Health System is the first energy consumer with great difference with other government administration sectors, such as Education or General Administration. Moreover, the energy bill has great impact on the sustainability of the public health care system. However, very few real data were available to characterize the energy demand on public buildings, which are usually the most intensive energy consumers, and efficiency indexes were usually obtained from simulation results. Nevertheless, thanks to the deployment of Smart Metering systems in the last years, it is possible to access to the true energy demand profiles of hundreds of these buildings.

In this paper, with three years historical monthly electrical energy consumption data from the health system of the region of Castilla y León in Spain -including hospitals, outpatient facilities, clinics and other medical institutions- and the application of data mining techniques, an end-use electrical energy analysis was conducted to cluster the building housing according to the energy consumption into several energy use intensity clusters and, then, an average value and a Reference Building Energy Index for each cluster is proposed. Thus, a true energy labeling of these buildings based on their distance to the Reference Building Energy Index is done and presented in georeferenced maps.

* Corresponding author. Tel.: +34-987-29-10-00, ext. 5391.
E-mail address: miguel.simon@unileon.es

Selection and peer-review under responsibility of the scientific committee of Technologies and Materials for Renewable Energy, Environment and Sustainability, TMREES18.
10.1016/j.egypro.2018.11.283

Keywords: Building Energy Index; Energy Labeling; Smart Metering; Clustering; Data mining.

1. Introduction

The European Union (EU) targets for climate and energy management for the year 2030 try to drive a new sustainable and renewable energy generation and consumption frame. This way, the main EU targets are focused in (i) the reduction of the energy consumption, (ii) the maximization of the energy efficiency, both for buildings and devices, and (iii) the development of new actions for energy savings, especially in the Public Administration.

Energy indexes seem to be a useful tool for the monitoring and supervision of the energy consumption and the greenhouse effect gases emission. Moreover, they provide information about trends in the historical energy consumption and become fundamental for energy planners and public administrators to develop efficient energy policies, from the local level to the national level.

Traditionally, energy efficiency labeling in the building sector has been carried out by energy simulations and efficiency labels are assigned in comparison with the results obtained by a reference building. This methodology works well with the thermal behavior of certain buildings and it has been demonstrated that accurate results can be obtained [1], but it seems not to be appropriate in large buildings, such as hospitals, as they involve several complex installations which many simulation systems cannot simulate accurately. However, very few energy labeling methods are applied to the power consumption of buildings – in most cases, apart from the electric heating and/or cooling systems, only illumination consumption is considered, as it affects significantly to the cooling or heating needs of a building-. This circumstance is because, (i) until nowadays, thermal (heating and/or cooling) demand was significantly greater (both in terms of magnitude and costs) than the power demand and, (ii) power needs are strongly related with the activity level, which experts do not come to an agreement to define it. However, with the deployment of nearly Zero Energy Buildings (nZEBs) and Positive Energy Buildings (PEBs), thermal demand in the building sector is decreasing dramatically, and power consumption is getting the main part of the energy bill [2]. Moreover, difficulties on the simulation of the power demand behavior of buildings can be overcome with access to real consumption measurements and the application of the so-called *"Big Data"* and *"data mining"* techniques. Thus, in this paper, as first approach to identify the energy efficiency of buildings and predict their power demand profiles, the authors propose a method to identify reference electrical energy consumption profiles (in terms of final energy use) by comparing several clustering techniques. As case study, the proposed methodology has been applied to the health system building stock of the Spanish region of Castilla y León, which includes more than 250 buildings. A proper identification of the power demand profiles will have great impact in the centralized purchasing of energy, Power Purchase Agreements (PPA), identify anomalies in the power consumption, find discrepancies with power energy audit reports, prioritize energy savings policies, among other applications [3]–[6].

This paper is structured into three more sections. The introduction section intends to introduce the reader into the origin of the need of the development of new Energy Building Indexes, motivated mainly by the latest updates of the EU Directives on energy efficiency and energy savings. Moreover, the health system building stock from the Castilla y León region is briefly depicted. Next section, entitled *"Materials and Methods"*, describes the methodology and the data characteristics. The third section shows the obtained results and conducts a brief discussion. Finally, in the last section, the authors present their main conclusions and propose future research lines.

Nomenclature

ACI Area Consumption Index.

Continued on next page.

BGF	Building Gross Floor.
EU	European Union.
EREN	Ente Regional de la Energía de Castilla y León (Public regional agency for energy management).
NF	Assignable area for main uses.
nZEB	Nearly Zero Energy Building.
OCC	Occupant.
OCI	Occupation Consumption Index.
OPTE	Optimización de la Tarifa Eléctrica (Electric consumption database and optimization tools).
PEB	Positive Energy Building.
PPA	Power Purchase Agreement.
RBEI	Reference Building Energy Index.

1.1. Main updates of the 2010/31/UE and 2012/27/UE Directives

The 30th May 2018, it was published the EU Directive 2018/844 [7] which updates Directive 2010/31/UE [8] on the energy performance of buildings and Directive 2012/27/UE [9] on energy efficiency, where the EU claims its commitment to develop a sustainable, competitive, secure and decarbonized energy system, at the time it remembers that the Energy Union and the Energy and Climate Policy Framework for 2030 establish the commitment, (i) to reduce greenhouse emissions by at least 40% by 2030 as compared with 1990, (ii) to increase the proportion of renewable energy consumed and, (iii) to make energy savings in accordance with the EU ambitions, as main goals. Moreover, the Union is committed to develop a fully decarbonized energy system by 2050, and to meet that goals, the EU concludes that Member States and investors need new measures to reach the long-term greenhouse gas emission goal. The presented roadmap pays special attention to the need of the decarbonisation of the building stock as fast as possible, because it is responsible for approximately 36% of all CO_2 emissions in the Union. This conclusion agrees with those from the 2015 Paris Agreement on climate change following the 21st Conference of the Parties to the United Nations Framework Convention on Climate Change (COP 21), boosting the EU efforts to decarbonize its building stock.

With the previously described framework in mind, the EU Commission finds the need to provide Member States and investors a clear vision to guide their policies and investment decisions, including national milestones and actions for energy efficiency to achieve in the short-term (2030), mid-term (2040) and long-term (2050). Thus, it results mandatory that the Member States specify the expected output of their long-term renovation strategies and monitor developments by setting domestic progress indicators, subject to national conditions and developments.

To achieve a highly energy efficient and decarbonized building stock and to ensure that the long-term renovation strategies deliver the necessary progress towards the transformation of existing buildings into nZEBs, or even to PEBs, in particular by an increment in deep renovations, it should be provided clear guidelines and, what is even more important, outline measurable, targeted actions [10].

Each long-term renovation strategy shall be in accordance with an applicable planning and encompass, among other conditions, (i) an overview of the national building stock, (ii) policies and actions to target all public buildings, and, (iii) an evidence-based estimate of expected energy savings and wider benefits, establishing measurable progress indicators. Moreover, databases for energy performance certificates shall allow data to be gathered on the measured or calculated energy consumption of the buildings covered, including at least the public buildings stock.

As final remark of the EU 2018/844 Directive regarding this work, it is pointed out that it will result mandatory to determine the energy performance of a building on the basis of calculated or actual energy use and it shall reflect typical energy use, not only for space heating, cooling or domestic hot water, but also for lightning and other electrical technical building systems [11].

1.2. The Health System in the Castilla y León region from the power consumption point of view

The Spanish region of Castilla y León, which location can be seen in Fig. 1, is one of the largest regions of the EU, covering an area of 94 223 km^2, representing the 18.6% of Spain, and accounts with 2 418 694 inhabitants in 2018

(5.39% of the Spanish population), being the 6th most populated region in Spain. Its GDP per capita is estimated in € 22 649 per person. This region is divided into 11 primary attention health areas, 14 specialized health areas and 14 administrative areas and, according to data from 2017, serves the medical needs of more than 2 354 500 patients, 258 000 of them younger than 13 years old (pediatric assistance) [12]. In this region, it can be found 7.81 health professionals per one thousand potential patients [12].

The building stock of the health system in Castilla y León includes hospitals, health centers (with and without emergencies), clinics, residences and administrative buildings and warehouses. As it can be seen in Table 1 and Fig. 2, hospitals account for a remarkable fraction of the electric energy consumption in the utility buildings sector (about the 80% of the total) while the rest contribute with just the 20% in a quite equally distribution. Nevertheless, this 20% of the total electric consumption of the building stock represents 25 GWh·yr^{-1}, which is, certainly not, a negligible energy consumption.

Fig.1. Location in Spain of the region of Castilla y León.

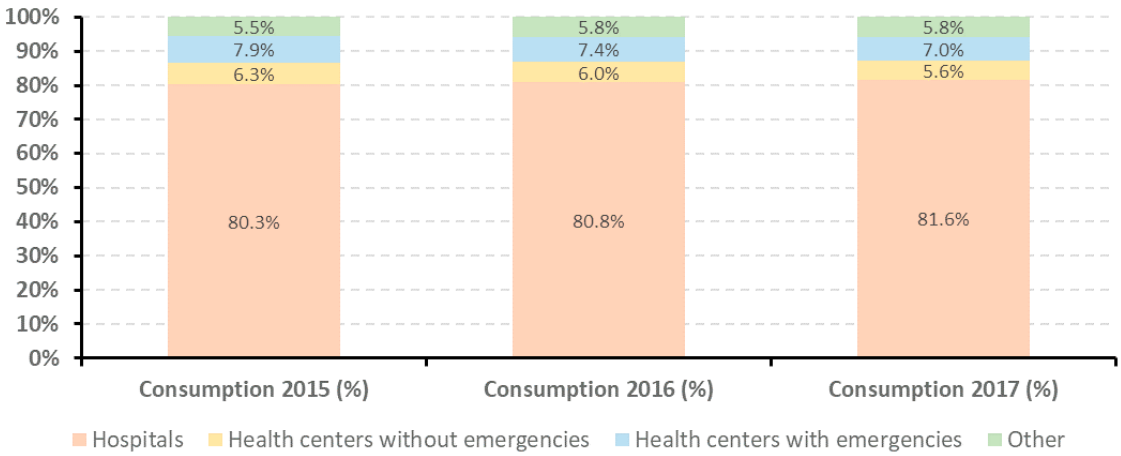


Fig.2. Evolution of the annual electric consumption of the building stock since 2015 until 2017.

It can be observed that, although the annual electric energy consumption distribution remains approximately the same, it can be seen an overall yearly increment. The overall increment between 2016 and 2015 was significantly higher than between 2017 and 2016, thanks to the important consumption decrement in health centers. Nevertheless, the highest differences are always observed in the hospitals consumption.

Finally, let's notice that the buildings classification seen in Table 1 is according to administrative purposes. One of the main targets of this paper is to provide the managers of health systems a tool to identify the reference buildings of their stock from an energetic point of view, independently of the administrative classification.

Table 1. Public Health System's building stock of Castilla y León. Relative increments regarding the previous year are represented in brackets.

Building type	Inventory	2015 consumption ($MWh \cdot yr^{-1}$)	2016 consumption ($MWh \cdot yr^{-1}$)	2017 consumption ($MWh \cdot yr^{-1}$)
Hospitals	27	100 690	108 877 (8.13%)	112 076 (2.94%)
Health centers with emergencies	176	9 932	9 948 (0.16%)	9 548 (-4.02%)
Health centers without emergencies	91	7 926	8 060 (1.69%)	7 627 (-5.37%)
Others	60	6 891	7 867 (14.17%)	7 023 (1.98%)
Total	354	125 439	134 753 (7.42%)	137 275 (1.87%)

1.3. Energy Indexes, average and reference values for buildings

When the energy efficiency of a particular building is getting analyzed, energy consumption per surface unit (square meter) is one of the most reliable indicators (Area Index). Depending mainly on the final use of the energy, gross surface, net surface, occupied surface or heated/cooled surface is considered, although occupied surface is used in most applications [13]. However, a surface indicator may not be representative in some cases, and the evaluation of the energy consumption per occupant (Occupation Index) can be a better option. This usually occurs with power efficiency studies where large buildings (great occupied area) can have low Area Index values, but can be poorly efficient if they have low occupation rates. Thus, in our study both occupied surface and occupation indexes have been considered to find clusters and Reference Building Energy Indexes (*RBEIs)*.

It should be taken into account that the isolated analysis of the energy indexes offers a particular point of view of the energetic behavior of a facility or building. Then, this sort of analysis must be conducted defining a levelized structure where building energy indexes can be aggregated and disaggregated accordingly to our analysis purposes. This aggregation capability can further explain changes along time and help to separate energy trends depending on its source: (i) activity level, (ii) structure, or (iii) energy intensity.

In this work, it has been considered the distinction pointed out in the standard VDI 3807 on the characteristic consumption values for buildings between demand and consumption characteristics [13]. Demand characteristics are calculated in accordance with the acknowledged rules of technology, using assumptions as to boundary conditions, standardized types of use and scenarios. On the other hand, consumption characteristics are determined on the basis of measured and corrected consumption values, if necessary. In this study, provided data are true consumption values from 3 years' monthly measurements. Thus, results will be expressed in terms of energy consumption instead of energy demand, although it could be applied to estimate future buildings' electrical energy demands.

Consumption analysis can be used during operation, e.g., as an initial value for the assessment of the energy consumption of a particular building, for comparison of buildings of the same type and use, for periodic assessments of the actual consumption and user behavior, as tool for management and controlling, etc., among many others.

The *RBEI* should include the characteristic consumption value, which is the generic term for the area-related characteristic of a building. It is determined from the energy consumption referred to the reference area of the building, during one year.

As reference quantities in this study, the occupied gross floor area or Building Gross Floor area (*BGF*), and the occupation (*O*), defined as the average number of beds in the case of hospitals and health centers, and average number of health professionals for the rest, have been used.

In this case, the reference area has been defined as the sum of all gross floor areas habitable of the building. In most cases, the habitable area is similar to the heatable area, which, according to VDI 3807 and DIN 277 standards, is calculated by subtracting major non-heatable gross floor areas from the building's gross floor area. The reference area of buildings in which only the full storeys are heated is identical with the storey area, which in general, can be taken from the building proposal. The floor areas are calculated from the outside dimensions of the full storeys, not taking into account any balconies, loggias, terraces and other building parts with no or but minor energy consumption. In absence of these data, according to the DIN 277 standard [14], the assignable area for main uses (*NF*) has been used or, if there no such value in this case either, the building's total gross floor area was used (in accordance with the German Energy Saving Ordinance, EnEV [15]). The VDI 3807 Part 2 standard, estimates the ratio *NF/BGF* in 85% for hospitals and health centers.

Thus, two reference indexes have been calculated and defined, the Area Consumption Index (*ACI*) and the Occupation Consumption Index (*OCI*), which expressions can be seen in equations (1) and (2), respectively. They can both be determined in monthly or annual basis.

$$ACI = \frac{E'}{A} = \frac{E'}{BGF}. \tag{1}$$

$$OCI = \frac{E'}{O}. \tag{2}$$

In equations (1) and (2), *E'* is the corrected energy in the time period, if necessary. Fig. 3 shows the mean annual *ACI* (heat map) and *OCI* (sized dots) representation according to the buildings location. It can be observed that, although in some cases high *ACI* values match with high *OCI* values, it can differ in many buildings. Figures in brackets show the quartiles' value.

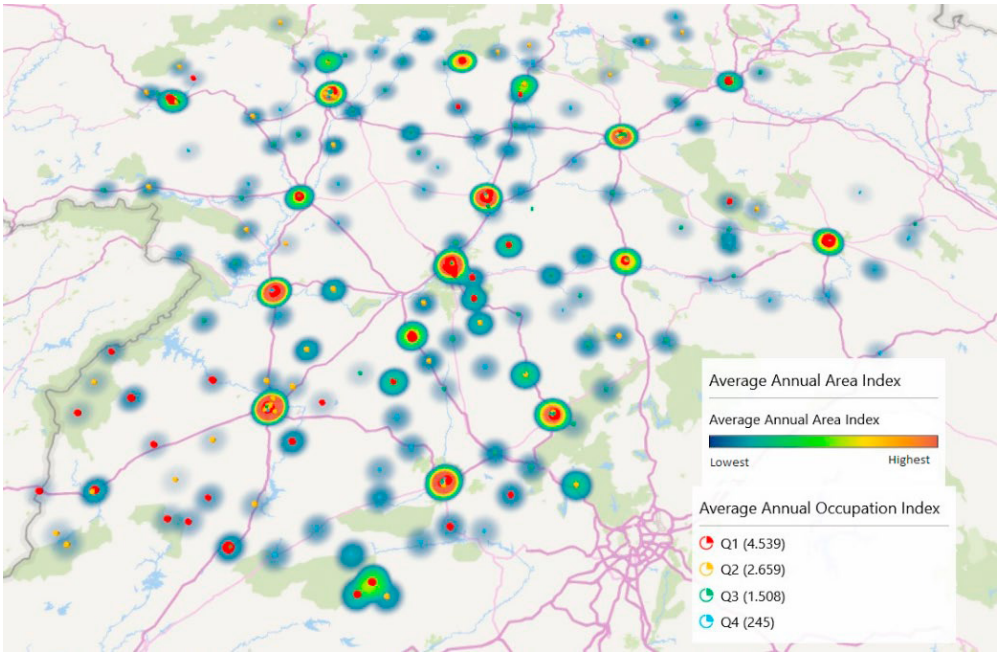


Fig.3. Geographical distribution of the annual *ACI* and *OCI*.

In contrast with heating or cooling energy, no outdoor-temperature effects corrections are usually needed for electrical energy, but time corrections if the measured period is not a complete year (365 days) must be done [16].

The development of reference values allows that the energy consumption of buildings can be evaluated approximately. If the characteristic consumption value of a particular building is higher than the average given for the building type, further analysis should be conducted. These analyses allow to identify existing shortcomings and potentials for improvement, in which case measures for a more efficient use of utilities should be taken. Moreover, the reference values given are characteristic values of consumption that occurred as real, favorable values in the buildings under consideration. The difference between the characteristic consumption value of a building and the relevant reference value allows to estimate an option for savings. For better comparability of the characteristic values, it is recommended to use a calendar year as a harmonized period for comparison.

Characteristic energy consumption value can be used for predicting the energy consumption of a large building inventory [17]–[19]. In the field of town and regional planning, for example, the can be used in estimating the demand of certain areas on the basis of the known building areas and types of building use, thus supporting the evaluation of various supply concepts [20].

It must be remembered that, as a consequence of changes in the building inventory, its equipment and the user behavior, the averages and reference values must be expected to shift with time. On the other hand, when comparing characteristic energy consumption values of buildings in other countries with averages and reference values in this work, the boundary conditions prevailing in those countries must be taken into account.

In accordance with already mentioned guide standards, such as VDI 3807 and DIN 277, the considered average value is not the arithmetic mean, but the modal value. The modal value is the value with the highest density of distribution, i.e. the most frequently occurring value of a distribution. The use of the modal avoids to yield in an excessively high reference value as frequency distributions of classified characteristic values of consumption are often oblique. In the case modal values cannot be calculated because of a small sample size, the median has been used as average reference characteristic value.

Finally, a class reference value of a characteristic energy index has been determined as the lower-quartile mean, which is the arithmetic mean of the lowest 25% of the characteristic values sorted in ascending order. This reference value (or *"good practices"* value) should not be used to describe the class (that is the average value), but shall be aimed for when implementing energy-savings measures or other investigations.

2. Materials and Methods

2.1. Database description

The Ente Regional de la Energía de Castilla y León (EREN) or *"Regional Department of Energy of Castilla y León"* has promoted an innovative application called OPTE (Optimización de la Tarifa Eléctrica or Power Tariff Optimization) which intends to homogenize the public energy contracts (both for fuels and electric energy) by helping local energy managers with the use of optimization tools. One of these tools, already deployed, collects the true power consumption of each Public Building registered in the platform. Thus, energy managers from SACYL (the Regional Health System) have registered, through the facility's CUPS (Universal Code for the Power Supply Point) each managed building, including hospitals, health centers, and administrative buildings.

Each building or installation in OPTE is characterized by a unique and invariant identifier, called IDOPTE. This IDOPTE allows the connection with other databases where other information can be provided, such as cadastral data, address, building manager, etc.

By default, OPTE organizes the buildings database according to an administrative criterion for accounting purposes and, although some pre-analysis tools are being implementing in the platform, no data analysis is provided apart from descriptive reports.

Hourly and monthly average power demand (provided by the Distribution System Operator) of each building, since 2015 is then available in the platform for downloading. Other installation data, such as type of energy contract, costs, pricing periods and so on are also provided with the power measurements. For this study, monthly data since January 2015 until December 2017 have been used.

2.2. Data filtering. Acceptance and exclusion rules

Initially, 354 buildings and facilities were available in the database, but a filtering process has been applied in order to discard errors and outliers which could disturb the results. Thus, the following exclusion rules have been applied:

- Those building references which have no available data on surface or occupation are discarded.
- Those building references which have an abnormal low power consumption (lower than the value of 100 $kWh \cdot occ^{-1} \cdot yr^{-1}$) are discarded.
- Those building references which have an abnormal high power consumption (greater than the value of 4000 $kWh \cdot occ^{-1} \cdot yr^{-1}$) are discarded.
- Those building references which have gaps or errors in the power measurements are discarded.
- Those building references which power measurements data breach normality of the data set are discarded.
- Those building references which power measurements data breach homoscedasticity of the data set are discarded.

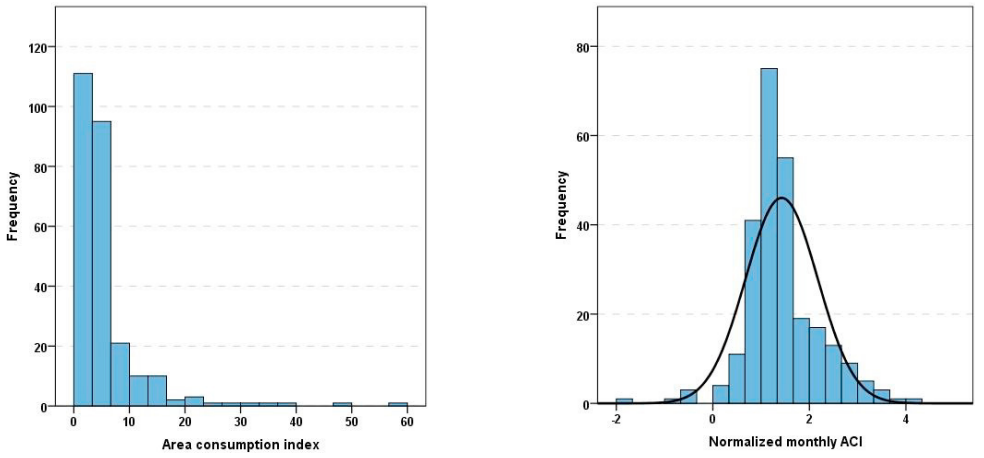
Thus, from 354 samples (building references), clustering techniques have been applied only to 259 samples, which implies a 26.84% of data rejection rate, which can be considered acceptable.

2.3. Data normalization and typification

In order to increase the clusterization algorithms performance, once that the energy indexes (area and occupation) have been calculated, they have been normalized and typified. Data have been normalized by applying the logarithmic transformation, while the typification has been conducted with the mean value subtraction and standard deviation division. Thus, final input data transformation for the clustering analysis can be seen in equation (3).

$$x' = \frac{\ln(x) - \overline{\ln(x)}}{\sigma[\ln(x)]}. \tag{3}$$

Fig. 4 shows a practical example for data samples before the application of the normalization and typification process. Sub-figure (a) shows clear absence of correlation with a normal distribution and values are in the range between 0 and 60, while sub-figure (b) shows data adjusted to a normal distribution centered in 1.



(a) (b)

Fig.4. Example of absolute frequency of monthly area indexes samples. (a) Raw values. (b) Normalized and typified values.

2.4. Clustering techniques

The data clustering techniques intend to find clusters from a dataset in a way that data items from the cluster are similar according to their parameters. These techniques constitute part of the kernel of the exploring data mining science and it has been widely applied in statistics. The clustering analysis cannot be defined as an algorithm itself but a bunch of them with many different orientations that can be applied to find clusters in a dataset. A more precise definition of the clustering is that it constitutes a multi-target optimization problem where a distance function, a density threshold and the number of clusters definition are involved. Moreover, the clustering analysis is not an automatic process, but an iterative one (interactive multi-target optimization) which implies a test and fail procedure [21].

There exist multiple clustering techniques and algorithms which can be classified into four main categories:

- Connectivity models: they are based on the distance analysis of the connections. Hierarchy methods are included in this category.
- Centroid models: where each group is represented by a vector of the mean values of the parameters (centroid). The most representative model in this category is the *k*-means model [3].
- Distribution models: groups are modeled by statistical distributions, such as the normal multivariate distribution (Expectation-maximization algorithm).
- Density models: groups are defined as dense regions connected in the data space (DBSCAN or OPTICS algorithms).

Moreover, the clustering can be hard (each member only belongs to one group) or soft (each member can belong to several groups simultaneously with different rate of belonging).

The most appropriate algorithm for clustering depends on the problem characteristics and, most times, it must be selected experimentally with the help of the researcher previous experience [4]. In this case, two clustering techniques types have been used and compared: hierarchy methods and non-hierarchy methods.

The hierarchy methods, or clustering based on connectivity, are based on the hypothesis that closer objects are more probably related that those which are far away. Thus, a distance function is mandatory in this sort of algorithms. Furthermore, these algorithms provide a hierarchy of groups which get fusion at the corresponding distance. This hierarchy structure is usually represented by a dendrogram, a 2D graph where the objects or samples are represented in one axis (usually the vertical one) and the distance is represented in the orthogonal axis. These algorithms work in a sequential form and the researcher can observe in each step the clusters division or association. However, these algorithms are sensitive with data noise and exponentially complex, which makes them not appropriate for large datasets [22]. In our case, this clustering algorithm has been applied just to estimate the number of clusters which can be significant in the dataset.

The clustering methods and the distance calculation find two main targets:

- To obtain clusters which individuals are the most similar, i.e. the mathematical distance between each individual and the centroid is the smaller as possible (intra-group distance).
- To obtain clusters which distances between centroids are the larger as possible, which means that the clusters are very different one from each other (inter-group distance).

From the hierarchy methods, in this study it has been applied the nearest neighbor algorithm and the Ward's clusterization algorithm, as the shape of the obtained dendrogram with the nearest neighbor algorithm applied first (many clusters of 1 single individual) suggest the Ward's clusterization could improve the results. This method is based on the hypothesis that in a clusters joint some information is lost. Thus, only those clusters which joint produces the minor information lost will be joined. This method uses Euclidean distances and helps to obtain small size clusters, avoiding the gravity effect from massive clusters [23].

On the other hand, the non-hierarchy methods, or methods based in the centroid, calculate the centroid of a cluster as a central vector. The centroid, which is not necessary that belongs to the sample, represents the cluster. One of the

most widely used algorithm of this category, because of its high performance and simplicity, is the *k*-means algorithm or Lloyd's algorithm [24]. It finds the *k* centroids of the *k* clusters and assigns members to each cluster according to their distance to the centroid. This definition constitutes a NP-hard optimization problem and thus, only approximations of the solution are feasible to compute. As it only finds local optimal values, the algorithm must be executed in an iterative way with random initial conditions. As main disadvantages of this algorithm is that, (i) it optimizes centroids and, thus, it can fail in the border definition of the clusters, and (ii) the number of clusters (*k*) must be set as initial condition and it is usually not known [3].

2.5. Analysis methodology

To properly apply the clusterization techniques described in the previous section, the following steps have been followed:

1. First, the dendrogram by the nearest neighbor algorithm has been obtained.
2. According to the dendrogram's shape, it is discussed if a more precise hierarchy clusterization technique is necessary to be applied. For all analyzed cases in this study, it was observed that the Ward's algorithm application would be desirable. Thus, the Ward's algorithm for clusterization is applied and the result is taken into account to observe the maximum number of clusters feasible for non-hierarchy methods.
3. The *k*-means non-hierarchy clusterization algorithm is evaluated for the dataset in an iterative way, introducing as the number of clusters from 1 to the previous result value in an iterative process.
4. The sum of intra-clusters' squares (error) and the sum of inter-cluster's squares (explained) are analyzed for each case characterized by the number of clusters to find. As the greatest number of clusters is defined, the best results (greatest explained value and lowest error value) will always be obtained, the relative improvement from the previous step is also studied. Then, a relative improvement lower than 10% has been chosen as stop criteria.
5. Once the optimum number of clusters is obtained, the average and reference (*RBEI*) values for each class are obtained (labeling). Moreover, graphs comparing the values for each class are drawn.
6. Finally, results are represented in 2D maps to observe the geographic distribution of the different building classes (mapping) according to the indicators.

3. Results and discussion

Table 2. Clusters' average values for annual ACI and OCI indicators (Ward's clusterization).

Cluster	Samples	Average ACI 2015 (kWh·m^{-2}·yr^{-1})	Average ACI 2016 (kWh·m^{-2}·yr^{-1})	Average ACI 2017 (kWh·m^{-2}·yr^{-1})	Average OCI 2015 (kWh·occ^{-1}·yr^{-1})	Average OCI 2016 (kWh·occ^{-1}·yr^{-1})	Average OCI 2017 (kWh·occ^{-1}·yr^{-1})
Class 1	63	6	6	4	1 334	1 375	1 113
Class 2	38	11	10	10	370	349	346
Class 3	6	28	27	26	4 142	4 018	3 820
Class 4	52	37	37	36	1 410	1 396	1 360
Class 5	100	148	151	143	10 897	10 899	10 747

3.1. Clustering results with hierarchy clusterization

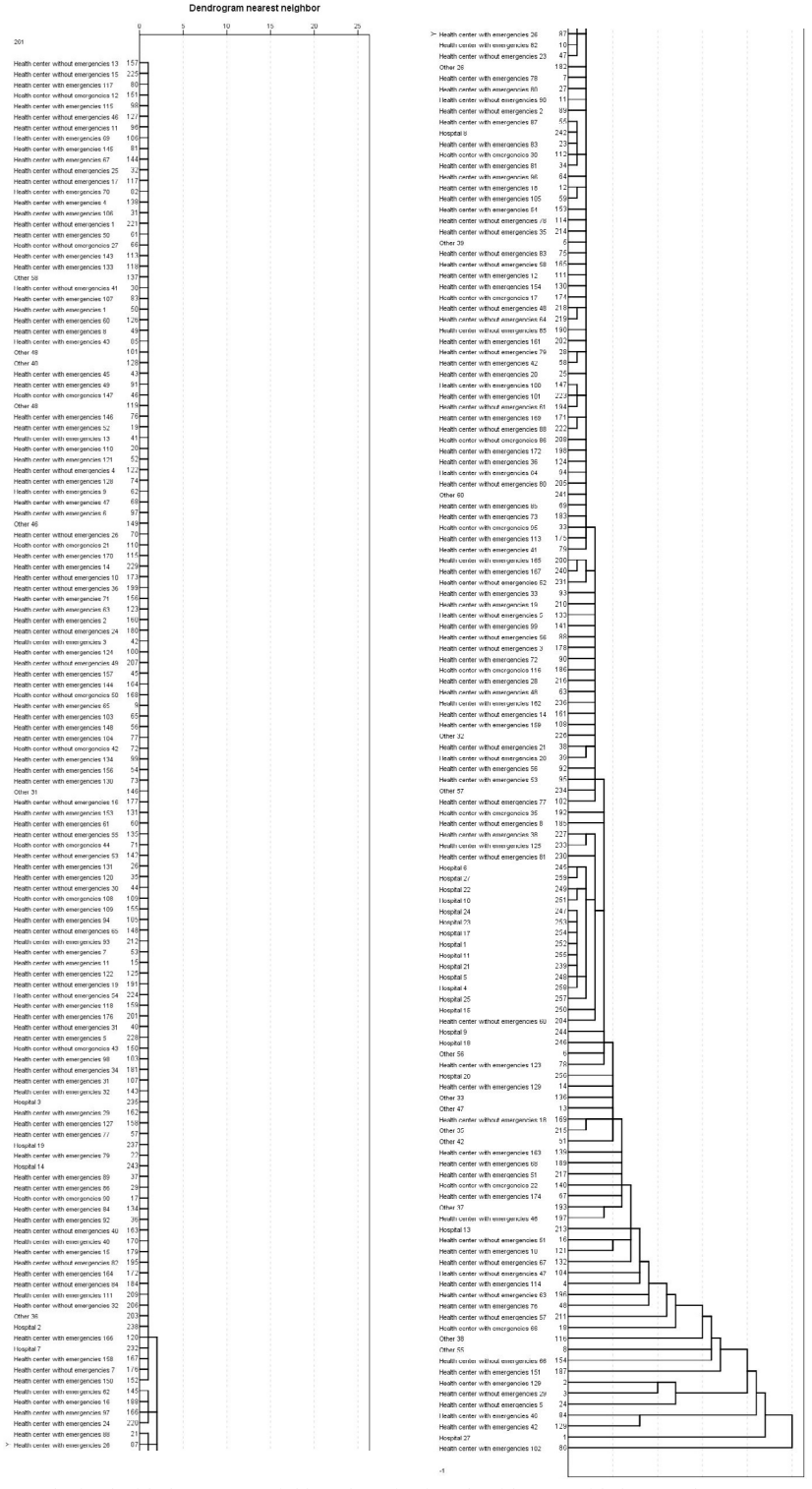


Fig. 5. Dendrogram obtained with the nearest neighbor clusterization algorithm (provided as supplementary material for details).

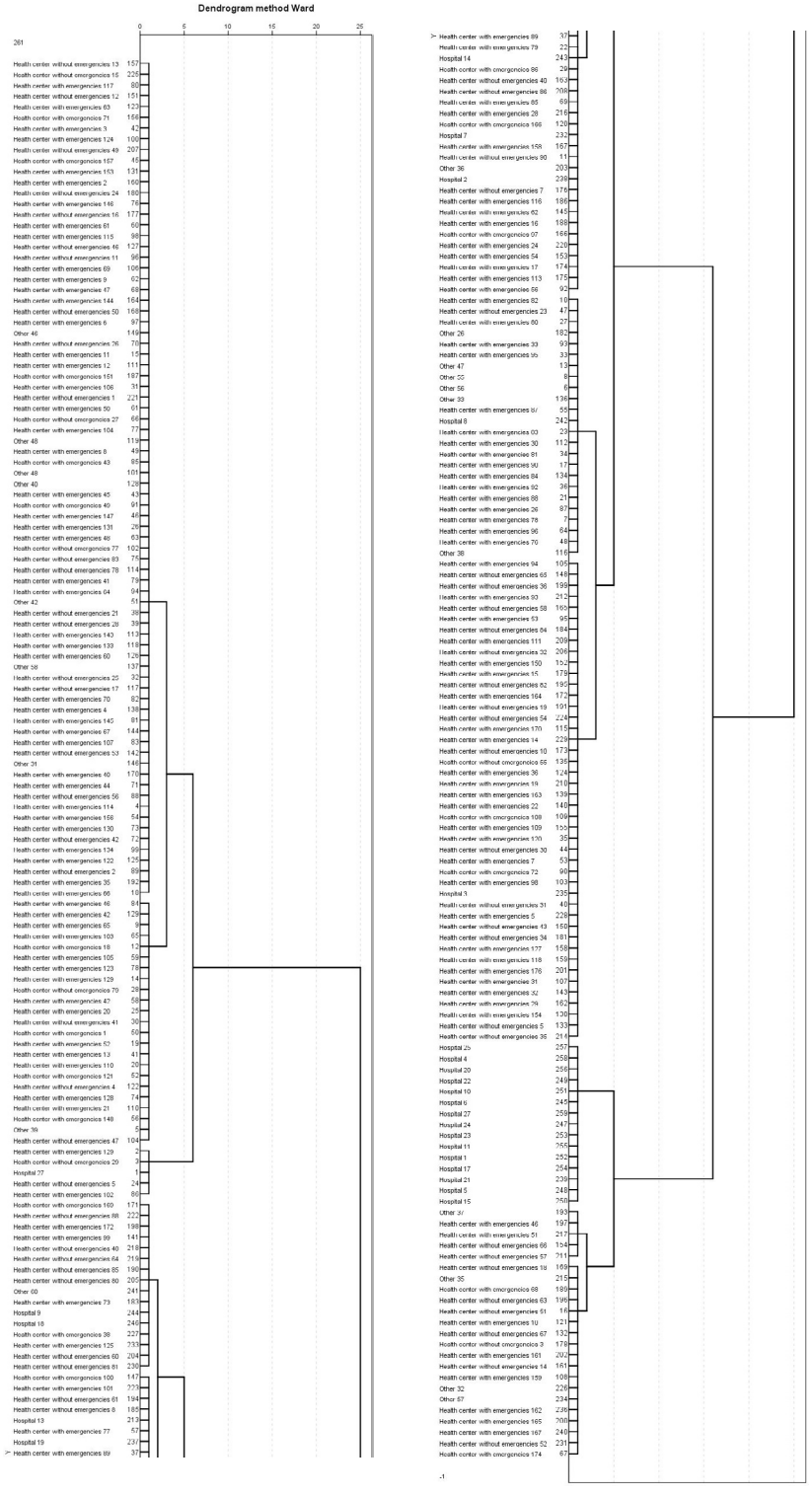


Fig. 6. Dendrogram obtained with the Ward's clusterization algorithm (provided as supplementary material for details).

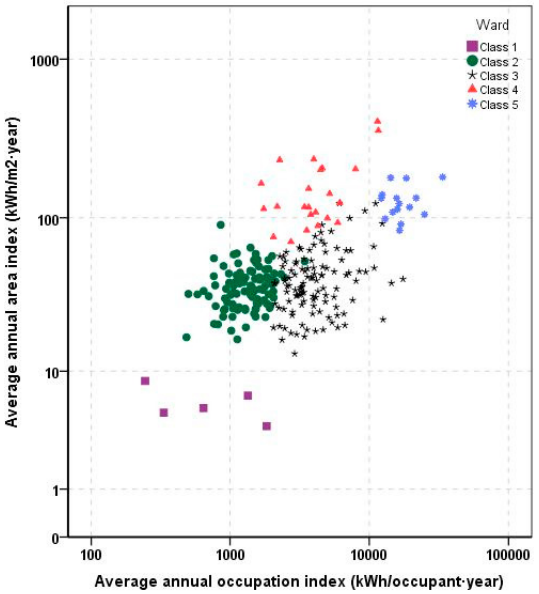


Fig. 7. Ward's clusterization result with 5 optimal clusters.

As it can be seen in Fig. 5, the obtained dendrogram by the nearest neighbor clusterization algorithm shows a stacked structure shape. Thus, the Ward's clusterization algorithm, which dendrogram can be seen in Fig. 6, seem to be more appropriate. In fact, as seen in Fig. 6, Ward's clusterization avoided the gravity effect and obtained smaller clusters. According to the distances in the Ward's dendrogram, 5 different clusters seem to be the optimal number of groups. Table 2 and Fig. 7 show the characteristics of the 5 clusters. It can be observed that the larger cluster is the fifth one, and the smaller, the third one. Moreover, annual *ACI* and *OCI* values for each class seem to be similar for the three years. Class 5 is characterized by the highest *ACI* and *OCI* values, followed by Class 3, while Class 2 seems to gather the lowest combination of *ACI* and *OCI* values (Class 1 has lower *ACI* values but significantly higher *OCI* values). Thus, 5 different reference buildings seem to be identified by this method.

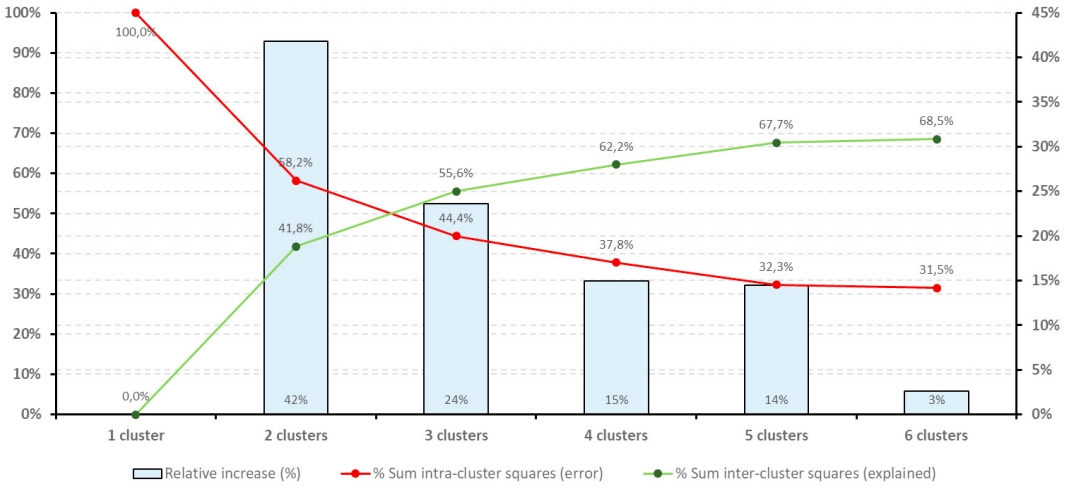
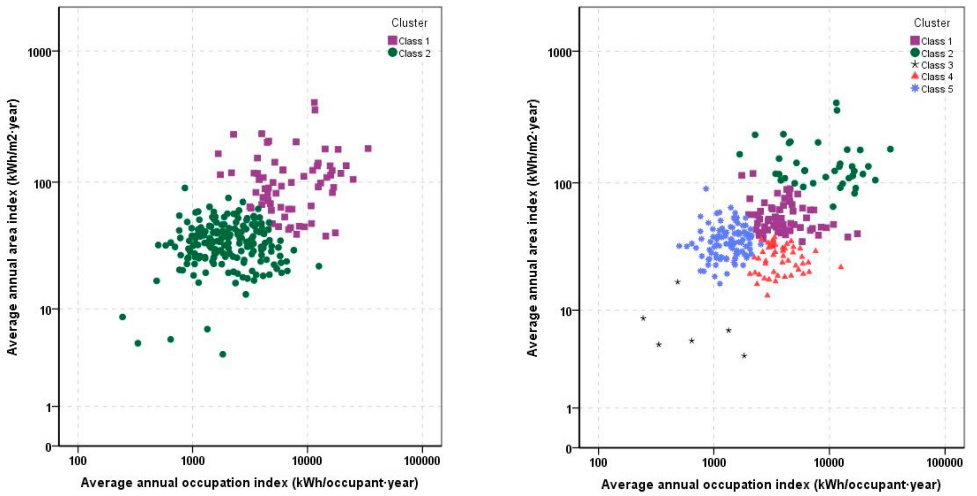


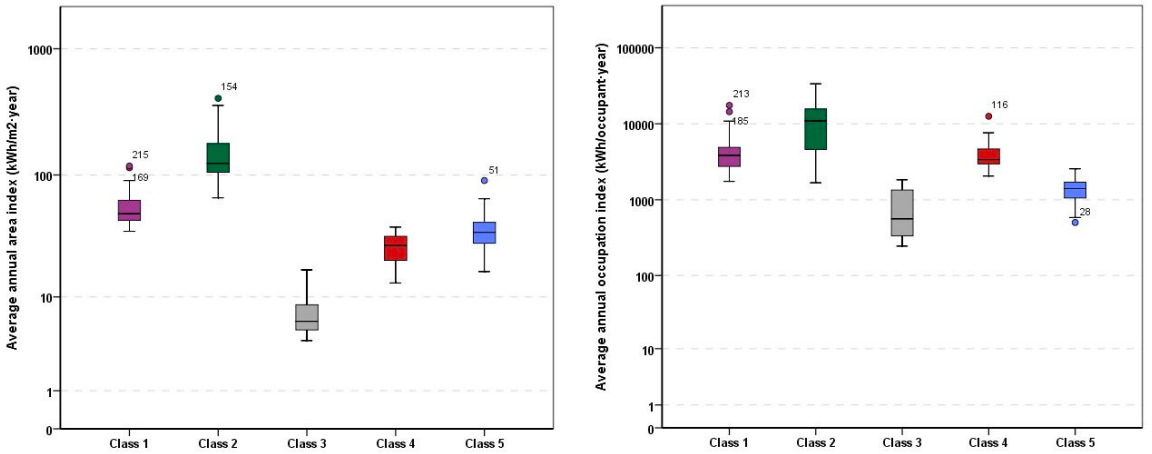
Fig. 8. k-means clusterization algorithm evaluation according to the input number of clusters.

3.2. Clustering results with non-hierarchy clusterization algorithms



(a) (b)

Fig. 9. *k*-means clusterization results graph for (a) two clusters and (b) five clusters.



(a) (b)

Fig. 10. Mean annual (a) *ACI* and (b) *OCI* values for each cluster.

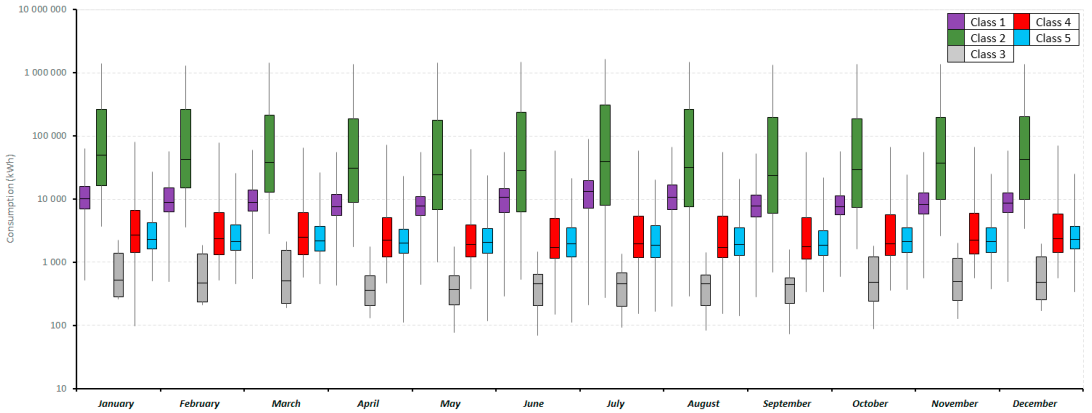


Fig. 11. Time distribution of electrical energy consumption for each class.

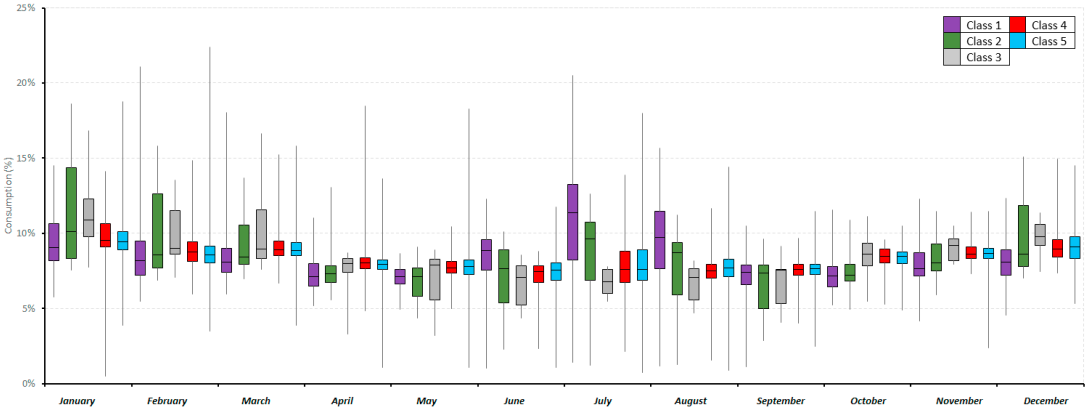


Fig. 12. Time distribution of the relative electrical energy consumption for each class.

Fig. 8 shows a graph where the clusterization error (sum of the intra-cluster squares), the explained rate (sum of the inter-cluster squares) and the relative increments are represented considering the clusterization results of the *k*-means algorithm (non-hierarchy method). It can be observed in this figure that the optimal number of cluster is 5. Although the highest the number of clusters, the lowest error, the relative improvement between 6 and 5 clusters is lower than 5%.

Sub-figures 9 (a) and 9 (b) show the results of the *k*-means algorithm application with two and five clusters, respectively. In the two clusters results, class 1 group the highest *ACI* and *OCI* samples, while class 2 gather the lowest *ACI* and *OCI* samples. However, with 5 clusters, more intermediate levels can be appreciated. Extreme values are associated with class 2 (very high intensive energy consumers) and class 3 (very low intensive energy consumers). These observations are supported by sub-figures 10 (a) and 10 (b), which show the distributions of the mean annual *ACI* and *OCI* for each cluster, respectively.

Finally, Figs. 11 and 12 show the time distributions of the electrical energy consumption for all clusters. Fig. 11 evaluates the total consumption value, while Fig. 12 shows the distribution of the monthly percentage consumption referred to the total annual electric energy consumption. It can be observed in Fig. 11 that there exist a clear distinction in the mean monthly consumption between classes, while in Fig. 12 it cannot be found significant differences in the monthly distribution between classes, with the exception of classes 1 and 2. Buildings classified in class 1 show a peak consumption in summer, while buildings from class 2 show more consumption in the winter period.

3.3. Comparison of the clusterization algorithm's results

Table 3 is structured in a similar manner than a confusion matrix where columns correspond with the classification from the application of the *k*-means algorithm, and rows correspond with the classification obtained from the Ward's algorithm calculation. First figure of each cell shows the number of samples, while the second figure shows the percentage. It can be seen that class 3 from the *k*-means algorithm corresponds with the class 1 of the Ward's method.

On the other hand, class 2 from the *k*-means algorithm corresponds with class 5 of the Ward's method. The rest of the classes show more *"confusion"*, specially class 3 from the Ward's method. Nevertheless, it is highly probably that class 2 from the Ward's method shares properties with the class 5 of the *k*-means algorithm, as the class 4 from the Ward's method is similar to the class 2 of the *k*-means algorithm.

3.4. Classes characterization: average and reference values

Table 4 shows the existing correlation between the *k*-means clusterization and the administrative classification of the buildings. As expected, the major part of the hospitals is clustered in class 2, which gathers the most energy intensive buildings. However, a 13% of the hospitals are clustered in class 1 and a 17% in class 4. On the other hand, health centers are classified mainly in clusters 1, 4 and 5. Very few differences can be observed between health centers

with and without emergencies. It can be observed that health centers with emergencies show a major proportion in class 4 than health centers without emergencies.

Table 3. k-means algorithm and Ward method results comparison.

		***k*-means algorithm**					
		Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Total**
Ward method	**Class 1**	0 / 0%	0 / 0%	5 / 100%	0 / 0%	0 / 0%	5
	Class 2	4 / 4%	0 / 0%	1 / 1%	0 / 0%	96 / 95%	101
	Class 3	53 / 46%	6 / 5%	0 / 0%	52 / 45%	4 / 3%	115
	Class 4	6 / 26%	17 / 74%	0 / 0%	0 / 0%	0 / 0%	23
	Class 5	0 / 0%	15 / 100%	0 / 0%	0 / 0%	0 / 0%	15

Table 4. Comparison between the administrative inventory and the *k*-means clusterization result.

		Inventory				
		Hospitals	**Health centres with emergencies**	**Health centres without emergencies**	**Others**	**Total**
***k*-means Algorithm**	**Class 1**	3 / 13%	31 / 22%	26 / 38%	3 / 13%	63
	Class 2	17 / 71%	11 / 8%	7 / 10%	3 / 13%	38
	Class 3	0 / 0%	3 / 2%	2 / 3%	1 / 4%	6
	Class 4	4 / 17%	34 / 24%	7 / 10%	7 / 30%	52
	Class 5	0 / 0%	65 / 45%	26 / 38%	9 / 39%	100

Table 5 collects the average (modal) and *RBEI* values for each class from the clusterization. Moreover, the standard deviation is also shown to see the indicators dispersion. The highest annual average ACI value corresponds to class 2, while the lowest corresponds to class 3. On the other hand, the highest annual average OCI value corresponds to class 4, while the lowest corresponds to class 3. This means that low energy intensity consumption buildings are both in area and occupation terms, while high energy intensity consumers can find discrepancies between area and occupation terms.

Finally, Fig. 13 shows the geographical distribution of the clustered buildings, from the two clusters analysis (sub-figure a) to the five clusters analysis (sub-figure d). Classes 1, 2 and 3 seems more concentrated on the regional capital cities, while classes 4 and 5 look more dispersed in the geography.

Table 5. Average and *RBEI* values for each class.

		Average annual ACI ($kWh \cdot m^{-2} \cdot yr^{-1}$)	**RBEI annual ACI** ($kWh \cdot m^{-2} \cdot yr^{-1}$)	**Standard deviation** ($kWh \cdot m^{-2} \cdot yr^{-1}$)	**Average annual OCI** ($kWh \cdot occ^{-1} \cdot yr^{-1}$)	**RBEI annual OCI** ($kWh \cdot occ^{-1} \cdot yr^{-1}$)	**Standard deviation** ($kWh \cdot occ^{-1} \cdot yr^{-1}$)
***k*-means algorithm**	**Class 1**	45.00	39.49	17.83	3 166.67	2 335.51	2 836.97
	Class 2	120.00	91.46	70.44	3 150	3 426.07	7 091.06
	Class 3	6.22	4.34	4.83	565.27	285.46	651.25
	Class 4	23.75	18.07	6.72	3 400	2 526.35	1 772.87
	Class 5	33.33	24.03	11.47	1 291.67	840.06	436.27

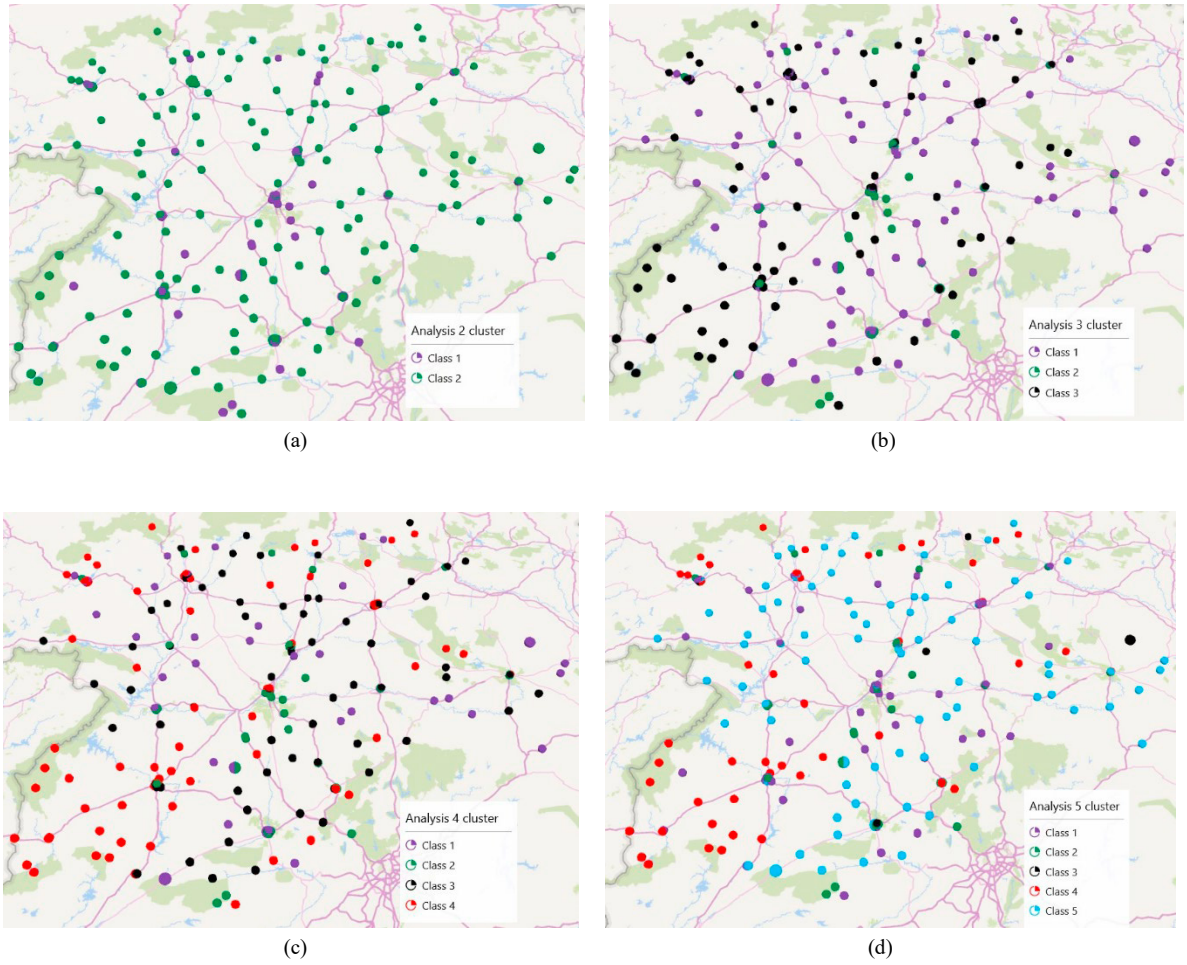


Fig. 13. Geographical distribution of the classified buildings for (a) two clusters, (b) three clusters, (c) four clusters and (d) five clusters.

4. Conclusions

The EU targets for climate and energy management for the year 2030 try to drive a new sustainable and renewable energy generation and consumption frame. Energy indexes seem to be a useful tool for the monitoring and supervision of the energy consumption and the greenhouse effect gases emission. Moreover, they provide information about trends in the historical energy consumption and become fundamental for energy planners and public administrators to develop efficient energy policies, from the local level to the national level.

In this paper a novel approach to find reference buildings by the application of clustering techniques is presented and applied to the buildings stock of the Castilla y León region in Spain. Results show a high correlation between the hierarchy method of the Ward's algorithm for clustering and the *k*-means algorithm (non-hierarchy method).

Moreover, the administrative classification traditionally conducted for this sort of buildings can conduct to incorrect energy analysis as this classification is not fully in accordance with the clusterization results. This fact specially affects to health centers (with and without emergency services), which, depending on their characteristics, can show different electric consumption behavior.

Furthermore, average values (calculated as the modal value), *RBEI*s and standard deviations have been calculated for each class. It is observed that the highest *ACI* values may not be associated with the highest *OCI* values, and vice versa.

Finally, it seems that there exist differences in the buildings classes distribution between urban and rural areas, which should be investigated in future research works. Moreover, although the *ACI* and the *OCI* energy indexes performed well, they show to be insufficient to cluster according to time distribution shifts on the energy consumption.

References

[1] A. L. Zorita, M. A. Fernández-Temprano, L.-A. García-Escudero, and O. Duque-Perez, "A statistical modeling approach to detect anomalies in energetic efficiency of buildings," *Energy Build.*, vol. 110, pp. 377–386, Jan. 2016.

[2] European Commission, *Communication from the Commission Europe 2020. A strategy for smart, sustainable and inclusive growth*. 2010.

[3] G. Tardioli, R. Kerrigan, M. Oates, J. O'Donnell, and 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.*, vol. 140, pp. 90–106, Aug. 2018.

[4] J. Yang *et al.*, "k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement," *Energy Build.*, vol. 146, pp. 27–37, Jul. 2017.

[5] C. Deb and S. E. Lee, "Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data," *Energy Build.*, vol. 159, pp. 228–245, Jan. 2018.

[6] F. Khayatian, L. Sarto, and G. Dall'O', "Building energy retrofit index for policy making and decision support at regional and national scales," *Appl. Energy*, vol. 206, pp. 1062–1075, Nov. 2017.

[7] European Union, *Directiva (UE) 2018/844 del Parlamento Europeo y del Consejo, de 30 de mayo de 2018, por la que se modifica la Directiva 2010/31/UE relativa a la eficiencia energética de los edificios y la Directiva 2012/27/UE relativa a la eficiencia energética.* 2018.

[8] European Union, *Directiva 2010/31/UE del Parlamento europeo y del Consejo de 19 de mayo de 2010 relativa a la eficiencia energética de los edificios (refundición)*. 2010.

[9] European Union, *Directiva 2012/27/UE del Parlamento europeo y del Consejo de 25 de octubre de 2012 relativa a la eficiencia energética, por la que se modifican las Directivas 2009/125/CE y 2010/30/UE, y por la que se derogan las Directivas 2004/8/CE y 2006/32/CE*. 2012.

[10] European Union, "Energy roadmap 2050," 2012.

[11] S. Papadopoulos, B. Bonczak, and C. E. Kontokosta, "Pattern recognition in building energy performance over time using energy benchmarking data," *Appl. Energy*, vol. 221, pp. 576–586, Jul. 2018.

[12] Junta de Castilla y León, "Recursos Sanitarios Públicos. Castilla y León 2017," Junta de Castilla y León, Castilla y León, Operación estadística 11013, 2018.

[13] VDI - Verein Deutscher Ingenieure, *Verbrauchskennwerte für Gebäude, Grunldlangen. Characteristic consumption values for buildings. Fundamentals. VDI 3807. Blatt 1/Part 1.* 2013.

[14] DIN, *Areas and volumes of buildings*. 2016, p. 14.

[15] International Energy Agency, *Energy Saving Ordinance*. 2002.

[16] VDI - Verein Deutscher Ingenieure, *Verbrauchskennwerte für Gebäude. Verbrauchskennwerte für t Heizenergie, Strom und Wasser. Characteristic consumption values for buildings. Characteristic heating-energy, electrical-energy and water consumption values. VDI 3807. Blatt 2/Part 2.* 2014.

[17] W. Chung, "Review of building energy-use performance benchmarking methodologies," *Appl. Energy*, vol. 88, no. 5, pp. 1470–1479, May 2011.

[18] W. Chung, "Using the fuzzy linear regression method to benchmark the energy efficiency of commercial buildings," *Appl. Energy*, vol. 95, pp. 45–49, Jul. 2012.

[19] A. B. R. González, J. J. V. Díaz, A. J. Caamaño, and M. R. Wilby, "Towards a universal energy efficiency index for buildings," *Energy Build.*, vol. 43, no. 4, pp. 980–987, Apr. 2011.

[20] S. Moghimi, F. Azizpour, S. Mat, C. H. Lim, E. Salleh, and K. Sopian, "Building energy index and end-use energy analysis in large-scale hospitals—case study in Malaysia," *Energy Effic.*, vol. 7, no. 2, pp. 243–256, Apr. 2014.

[21] X. Gao and A. Malkawi, "A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm," *Energy Build.*, vol. 84, pp. 607–616, Dec. 2014.

[22] D. Hsu, "Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data," *Appl. Energy*, vol. 160, pp. 153–163, Dec. 2015.

[23] F. Murtagh and P. Legendre, *Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm*. 2011.

[24] F. Wang *et al.*, "Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns," *Energy Convers. Manag.*, vol. 171, pp. 839–854, Sep. 2018.