POLITECNICO DI TORINO
Repository ISTITUZIONALE

Discovering Knowledge from a Residential Building Stock through Data Mining Analysis for Engineering Sustainability

Publisher copyright

(Article begins on next page)

04 August 2020

7th International Conference on Sustainability in Energy and Buildings

Discovering knowledge from a residential building stock through data mining analysis for engineering sustainability

Alfonso Capozzoli[a,*], Daniele Grassi[a], Marco Savino Piscitelli[a], Gianluca Serale[a]

[a]Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino 10129, Italy

Abstract

In this paper, a dataset of 92,906 dwellings was analysed adopting data mining techniques for the classification of heating and domestic hot water primary energy demand and for the evaluation of the most influencing factors. The sample was classified in three energy demand categorical variables (Low, Medium, High) considering different geometrical and physical attributes. The output of the model made it possible to set reference threshold values among the physical variables. Moreover, high energy demand dwellings were analysed in depth using a k-means algorithm in order to evaluate the design variables which need to be considered in a refurbishment process.

1. Introduction

The application of energy efficiency and sustainable green design features in new and existing buildings has become a crucial issue in recent years for building owners, designers, contractors, and facility managers [1,2]. An integrated whole building process throughout the entire project development process leads building designers to generate a large amount of data during energy simulations. These data need to be analysed in order to provide useful information for designers and authority planners, aimed at identifying the major causes of high energy consumptions and reference values to drive a building sustainability design approach. The aim of this research is to develop a tool that can help project teams and public authorities to evaluate and identify useful patterns in large data building

---

* Corresponding author. Tel.: +39 011 090 4413; fax: +39 011 090 6329.
  E-mail address: alfonso.capozzoli@polito.it

stocks. To this purpose, data mining techniques - which are data analysis processes combining different techniques from machine learning, pattern recognition, statistics to automatically extract concepts, interrelationships and patterns of interest from large datasets – were used in this paper. By applying data mining techniques in the analysis of a residential building stock, the identification of useful and previously unknown patterns can be identified [3, 4].

In this paper, the Primary Energy Demand (PED) - calculated according to the methodology proposed by EN ISO 13790 for space heating and Domestic Hot Water production (DHW) - was analysed for a database consisting of several residential dwellings. The analysis of dataset was retrieved from the Italian energy certificate database which contains information on envelope, technical plant features and PED calculated in "Standard Rating" conditions for each dwelling [5]. The "Standard Rating" approach could produce results for PED also very far from actual energy requests, because of standard assumptions regarding occupant behavior and ventilation are taken into consideration [6]. However, since a large dataset was analysed in this paper, the potential information that can be extracted in relation to the main patterns driving the PED can be considered consistent.

A pre-processing analysis in the first part was carried out in order to clean the dataset by removing outliers. Afterwards, a data transformation analysis was performed introducing criteria for labelling each building as having a "High", "Medium" or "Low" PED. The CART Method (Classification And Regression Tree), which consists in a multi-stage decision-making process to classify the observations in a finite number of classes, was implemented. The output of the model is a flow-chart subdividing the observations into homogeneous subsets [7] according to respect response, represented in our model by primary heating and domestic hot water energy demand. The tree was built considering the some important variables influencing the PED (aspect ratio, opaque and transparent envelope average U-values and global efficiency of the heating system). The classification process made it possible to introduce a set of decision rules in order to outline the splitting criteria. The outcome of this process consists in useful information that helps recognize the patterns which drive the evaluation of the energy performance of buildings analyzed. Furthermore, a detailed analysis was performed using k-means algorithm on the "High" consumption dwellings. This kind of analysis made it possible to divide the "High" consumption samples into similar groups. A benchmark value was evaluated to find an archetypal dwelling and some useful information was retrieved regarding the variables that need to be improved in order to obtain a lower PED for each cluster.

2. Materials and methods

2.1. Construction of the dataset

The data analysed in the present work was retrieved from a database of energy certificates for several dwellings. The dataset provide the annual PED for heating and domestic hot water consumption in 92,906 dwellings situated in the Piedmont region (Northern Italy) and included in the building types "multi-family houses" and "apartment blocks" of the Tabula project [8]. The PED was calculated using the "Standard Rating" methodology suggested in technical standard EN ISO 13790 and considering energy needs for DHW production and space heating. The DHW energy demand was calculated considering standard values referring to floor area, while the heating energy demand was evaluated considering building energy balance. The modelling of the building geometry considers real shapes (including complex, shapes) and self or over shading of other buildings. The quasi steady-state calculation method is based on the monthly balance of heat losses (transmission and ventilation) and heat gains (solar and internal) evaluated in monthly average conditions. Transmission heat losses were estimated taking into consideration opaque and transparent surfaces and as well as the thermal bridging effect. In "Standard Rating", parametrical values depending on floor area or heated net volume are taken into consideration when evaluating the ventilation rate and internal heat gains. The dynamic effects on the net heating energy demand are taken into account by introducing the dynamic parameters utilization factor and an adjustment of the set-point temperature for intermittent heating/cooling or set-back. These parameters depend on the thermal inertia of the building, on the ratio of heat gains to heat losses and on the occupancy/system management schedules. The annual PED is calculated from the net energy demand through different system efficiencies (heat emission, control of the internal temperature, heat distribution, heat generation) considering the thermal losses in the various sub-systems related to both heating and DHW. For the heating season, the average system efficiency represents the ratio between the annual net energy and the annual PED for heating and hot water production. The PED includes also the electrical energy demand of auxiliaries systems.

2.2. Description of the dataset

The dataset is composed by dwellings situated in the Piedmont region, in North Western of Italy. Given that, this region has a latitude of almost 45° and, according to the Köppen–Geiger climate classification system, it is situated almost within the entirely CFA (humid subtropical climate) zone, with some mountain areas situated in DFB (humid continental climate) zone. The standard heating season is 183 days long. In particular, the sample of dwellings analyzed are located in a climate with conventional Degree Days ranging between 2422 and 5165. However, the PED related to each dwelling was previously normalized referring to the 2617 DD of the city of Turin.

A frequency distribution analysis of the geometrical features of the samples, reveals that 44 % of the dataset is composed of dwellings with a floor area ranging between 60 m^2 and 90 m^2, 37 % ranging between 30 m^2 and 60 m^2, 15 % ranging between 90 m^2 and 120 m^2 and the remaining 4 % had other dimensions. Since the dataset is very large, the previous analysis could be representative of the typical dimensions of dwellings in Italy. Considering the construction periods, three different clusters were highlighted. The first one includes 38 % of the dataset and it is composed of dwellings built before 1960. In general, their physical characteristics are very poor and an energy refurbishment should be implemented. The second set considers the samples built between 1960 and 2005 while new apartments built within the last decade are included in the third cluster construction period. Respectively they refer to 58 % and 4 % of the dataset. In order to correctly categorize the dataset from an energetic point of view different input variables need to be considered. The dataset collects information relating to: year of construction and last refurbishment, floor area, heated volume, heat transfer surface, aspect ratio, average U-value of the opaque and transparent envelope, subsystem efficiencies of the heating plant, global efficiency of the heating and domestic hot water systems, boiler size and Italian energy classification. Through a sensitivity analysis the variables selected for the data mining analysis are listed below:

- aspect ratio (ratio of heat transfer surface on heated volume) [m^{-1}];
- average U-value of the opaque envelope [W/m^2K];
- average U-value of the windows [W/m^2K];
- global efficiency of the heating and domestic hot water systems [-].

Depending on the building shape the aspect ratio determines how large the surface exposed to the external environment is and consequently provides information on the heat gain and loss through the building envelope. The average U-value influences heat losses by transmission while global efficiency provides information on the quality of the plant system. In general, more recently built apartments are characterized by better technological solutions. Fig. 1 illustrates the frequency distribution of these variables according to the different building construction periods. From these charts, the technological improvements changing the performance of buildings in the last decade can be deduced. In particular, the building stock has reached U-values that are significantly below 0.8 and the majority of them present a global efficiency of the heating system of over 80 %. Meanwhile, for older buildings performed poorly.
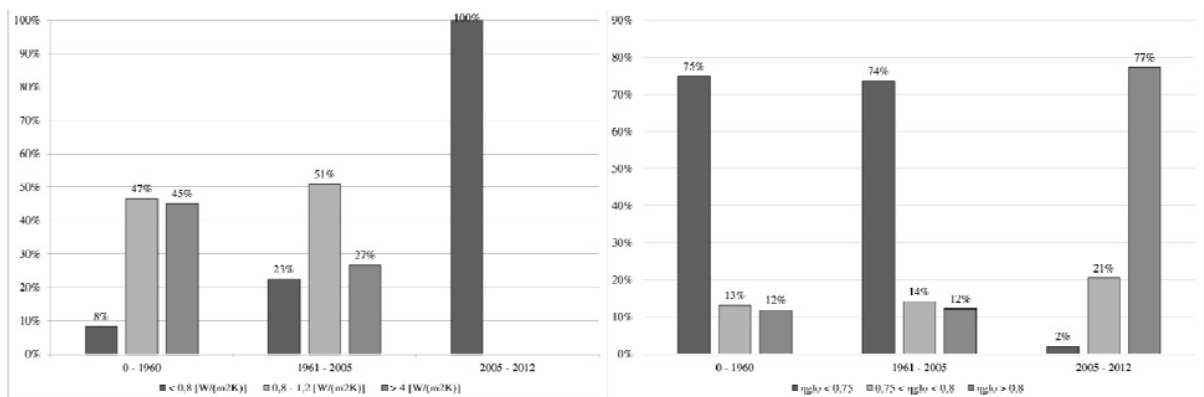
Fig. 1. (a) opaque envelope average U-value; (b) global efficiency of the heating system.
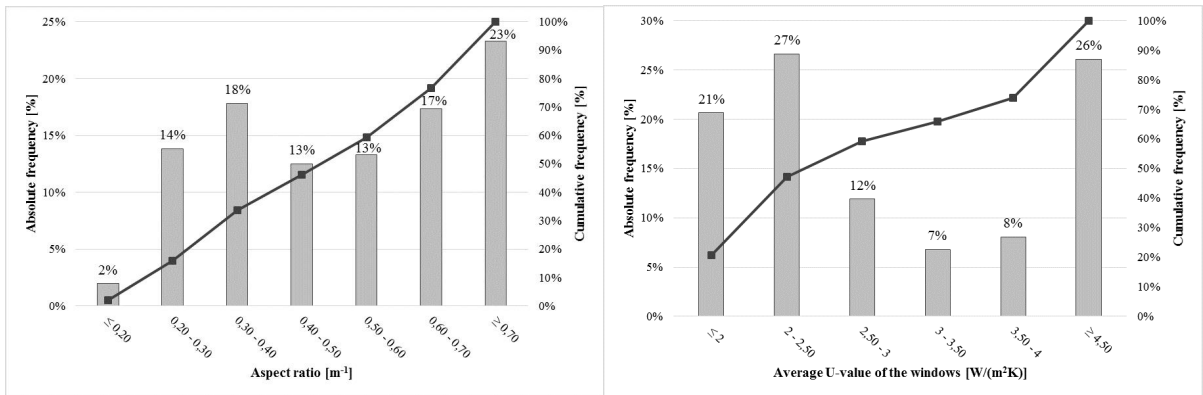

Fig. 2. (a) Aspect ratio frequency distribution; (b) average U-value of the windows frequency distribution.

2.3. Pre-processing analysis

A pre-processing analysis to identify the outliers is required before a classification model is created. An observation is an outlier when it does not follow the pattern of other data of the sample and appears to be inconsistent with the remainder of the rest of dataset. An analysis aimed at removing dataset outliers was performed. Moreover, the data were normalised on the floor area of each apartment, obtaining a normalised Primary Energy Demand (nPED). The average nPED of the dataset is 214.22 kWh/m^2, while the median is 205.54 kWh/m^2. Fig. 3 reports the frequency distribution of nPED.

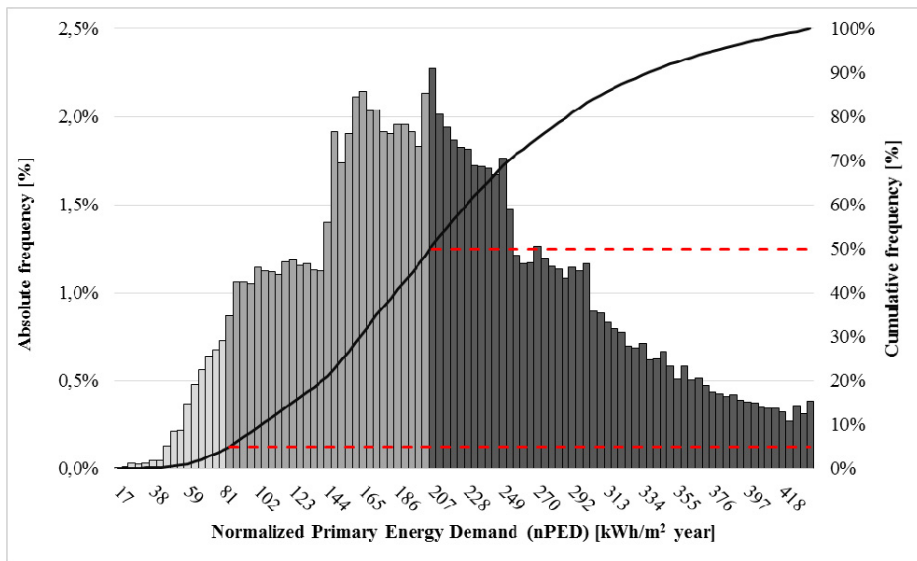Fig. 3. nPED frequency distribution, categories were highlighted with different grey pattern.

A data transformation analysis was performed introducing criteria for labelling each dwelling as "High", "Medium" or "Low" nPED. This data transformation is necessary for the construction of the classification tree. The selected thresholds between categories were set to 82 kWh/m^2 and 205.54 kWh/m^2 for dividing "Low" – "Medium"

and "Medium" – "High" dwellings respectively. The median of the dataset was used as a limit value for splitting "Medium" from "High" PED dwellings. For this reason, dwellings with a PED higher than 205.54 kWh/m^2 were classified as "High". In the Piedmont region, residential buildings with an energy demand lower than 82 kWh/m^2 are considered as high energy efficient buildings (energy class label A+, A and B). In this paper, the same criterion was adopted. The samples classified as "Low" represent 5 % of the dataset. This value could be used for a further generalisation of threshold value evaluation: in a generic dataset labelled "Low" samples are usually included in the 5th percentile of nPED. Moreover, through a normalization of nPED considering Turin DD, values are not depended on external climate conditions and which are adoptable also in other climate regions, were obtained (Table 1).

Table 1. nPED categories.

Consumption category	nPED per square meter [kWh/m^2]	nPEDdd per square meter and DD [kWh/DD·m^2]	Percentile
"Low"	$0 \leq nPED \leq 82$	$0 \leq nPEDdd \leq 3.13 \cdot 10^{-2}$	1-5
"Medium"	$82 \leq nPED \leq 205.54$	$3.13 \cdot 10^{-2} \leq nPEDdd \leq 7.85 \cdot 10^{-2}$	6-50
"High"	$nPED \geq 205.54$	$nPEDdd \geq 7.85 \cdot 10^{-2}$	51-100

2.4. Machine learning methods for dataset classification

In recent years, the techniques of machine learning, data mining and knowledge discovery in database were successfully applied on the PED data. Machine learning is the common term for supervised learning methods and originates from artificial intelligence [9]. In this scope, pattern recognition is a sub-area of machine learning and consist in the analysis of patterns within the data in order to identify a correct classification. The aim of pattern recognition is to learn a classifier data (patterns) based on prior knowledge or statistical information extracted from the pattern. In general, these classification algorithms are groups of measurements or observations, defining points in an appropriate multidimensional space. The common classification techniques are: ID3, C4.5 and CART. In this study using an open-source data mining software (RapidMiner), a CART algorithm was developed. This technique produces only a binary split (considering all 2k-1 ways of creating a binary partition of k attribute values) beginning with the Root Node, that contains the whole learning sample, and splitting each subsequent Parent Node into two Child Nodes. The split is an iteratively process that splits the dataset into sub-classes. The best way to divide the record depends on the type of measure chosen. This measure is defined in terms of the record's class distribution before and after splitting. In this work, the Gini index was used as a degree of impurity of each node. It is defined as Equation 1 where p(i|t) is the proportion of units in node t that belong to the j-th class of the response variable Y:

$$Gini(t) = 1 - \sum_{i=0}^{c-1} \left[p(i \mid t) \right]^2 \tag{1}$$

$$impurity_{SPLIT} = \left\{ Gini(t) - \left[Gini(t_l) \cdot p(t_l) + Gini(t_r) \cdot p(t_r) \right] \right\} \tag{2}$$

Equation 2 is used to evaluate the total impurity of the split. Where Gini(t) is the Parent Node impurity, Gini(t$_l$) is the impurity of the left node, p(t$_l$) is the proportion of units contained in the left node, Gini(t$_r$) is the impurity of the right node, while p(t$_r$) is the proportion of units contained in the right node. The best split s* is the one that maximizes the decrease in impurities. The statistical performance of each classification algorithm has to be evaluated in order to apply it to a new dataset. The k-fold cross-validation is the method used in this paper to evaluate the decision tree. This approach segments the data into *k* equal-sized partitions. This procedure is repeated *k* times so that each partition is used for testing exactly once. The total error is obtained through the sum of the errors for all *k* runs. For a further investigation of a determined group of samples, a cluster analysis was performed. This is an algorithm that allows for objects with similar characteristics to be grouped together into clusters. In particular, each cluster captures the natural structure of the data. Since the data are located in an n-dimensional space, the similarities according to distance-based metrics were evaluated. In this study, the Euclidian Distance was used in order to apply the k-Means algorithm correctly. This process requires as an input parameter the number *k* of

partitions. The optimal number of partitions (k) was valued using the minimization of Davies-Bouldin index, as the internal validation method. However, the output of this analysis is a just useful starting point for understanding the energetics features of apartment label as "High" nPED.

3. Results

A classification tree was built based on the most important variables influencing the nPED (aspect ratio, opaque and transparent envelope average U-values and global efficiency of the heating and domestic hot water systems). The classification process involved the introduction of a set of decision rules for the characterization of the splitting criteria.

Table 2. Confusion Matrix.

		Classified			
		Low	Medium	High	Correct
Real	"Low"	**3,188**	1,327	0	70.6%
	"Medium"	232	**33,151**	8,440	79.3%
	"High"	0	5131	**41,437**	89.0%
	Percentage				**83.7%**

Considering the four input attributes (aspect ratio, U-value opaque envelope and windows and global efficiency) a classification tree model was developed in order to predict the selected categorical variables ("Low", "Medium", "High"). A pruning analysis was carried out to remove the Leaf Nodes which did not improve the classification process. In this approach, by setting the minimum number of cases in Parent and Child Nodes – 1,000 and 800 cases respectively – and the maximum decrease in impurities of each split – ImpuritySPLIT = 0.001 (Eq. 4) –   the decision tree was initially developed to its maximum size. Finally, each Leaf Node with an error rate higher than 25 % was removed. Each Leaf Node in the final tree contains at least 1 % sample of the total and has a minimum accuracy of 75 %. To estimate statistical performance of the learning process, the number of validation k equal to 15 (cross-validation) was set. In Table 2 the Confusion Matrix is reported, illustrating for each class how instances from a specific class received various classifications. The columns show the real categorical label attribute whereas the rows illustrate the label attribute given by the classification. The number of correctly classified cases appear diagonally in Table 2. Considering, for example, the final class "Low" the confusion matrix indicates that 3,188 dwellings are correctly classified while 1,327 dwellings are misclassified as "Medium". The table also shows that there are always fewer incorrect classified cases than correct cases. The last row, shows that 83.7 % of all training records are correctly classified as "Low" – "Medium" – "High" nPED. This value indicate a good accuracy of the CART algorithm and its usability to new classification dataset.

Table 3. nPED categories and classification criteria.

Leaf	Variables				Amount	
Low	Uo≤0.45	-	η>0.84	-	4,275	4.6%
Medium	0.45≤Uo≤0.84	-	η>0.84	-	1,600	1.7%
	Uo≤0.84	-	η≤0.84	-	15,741	16.9%
	Uo>0.84	-	η>0.83	S/V>0.46	2,171	2.3%
	Uo>0.84	Uw≤4.00	-	S/V≤0.46	15,553	16.7%
	Uo>0.84	Uw>4.00	η>0.66	S/V≤0.46	5,416	5.8%
High	Uo>0.84	-	η≤0.83	S/V>0.46	39,468	42.5%
	Uo>0.84	Uw>4.00	η≤0.66	S/V≤0.46	8,682	9.3%

The algorithm can be translated into a set of decision rules, which have the following form: *if antecedent conditions, then consequent conditions*. In Table 3 the results of the CART algorithm are presented in terms of decision rules, starting from the Root Node and following all the possible ways of reaching each Leaf Node. The first column titled "Leaf" shows the final nodes of the tree which classify the nPED. The second column shows the rules that have to be respected in order to classify a dwellings in categorical variable, considering the conditions in different rows. The third column indicates the amount of samples included in a final node and their percentage on the total dataset.

The boxes in Fig. 4 represent the different nodes of the classification tree. The first node is the Root Node which considers the whole dataset of 92,906 samples. The Leaf Nodes report the final category of nPED in which the samples are classified. Furthermore, in each node the number of samples split and their percentage of the total are also reported. When the node is not a Leaf Node, the logic condition for the following split is marked in the third row. In such a case, Branch Y (yes) is to be followed, otherwise Branch N (no) should be followed.
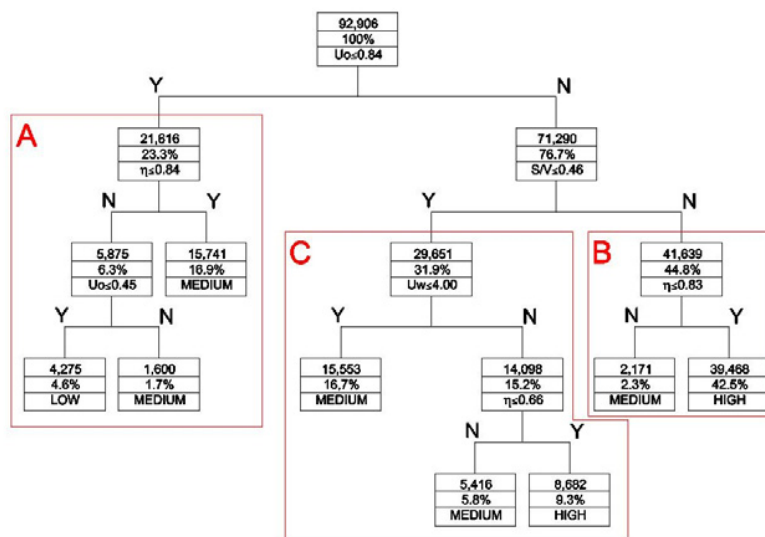

Fig. 4. Classification tree obtained using the CART algorithm.

4. Discussion

4.1. Critical analysis of the classification tree split variables

Useful information can be obtained from a decision tree model, for example, it helps to understand the energy consumption pattern of a dwelling and the way to optimize the building design. Therefore, by examining the decision rules, the significant factors influencing nPED profiles can be identified [10]. The first split is driven by the variable that most influences the nPED. As shown in Fig. 4 the average U-value of the opaque envelope is the first split variable of the classification model. In particular, with U-values lower than 0.84 W/m^2K all the dwellings are classified as "Medium" or "Low". This part of the tree is highlighted in Fig. 4 with the area marked as A. In this area the dwellings classified as Low nPED ($\leq$ 82 kWh/m^2) can be divided from dwellings with "Medium" nPED. Comparing the threshold U-value with the ones reported in Fig. 1(a) it is clear that each apartment built after 2005 is included in area A. In particular, following the rules listed in Table 2, the "Low" samples are characterized by average U-values of the opaque envelope lower than 0.45 W/m^2K and a global energy efficiency higher than 0.84.

If the average U-value of the opaque envelope is higher than 0.84 W/m^2K the Branch N should be followed after the first split. The Child Node divides into two categories the data taking into account the apartment aspect ratio.

This second splitting highlights that for dwellings with higher opaque U-value, the principal variable affecting the nPED is the aspect ratio. If the aspect ratio is higher than 0.46 the B area is defined. In general, the samples in this area are mainly classified in the "High" energy demand class. Only a small percentage with global efficiency of the heating system higher than 0.83 belongs to the "Medium" nPED category.

Finally, an aspect ratio lower than 0.46 leads to C area. Once again, the dwellings included in this area belong to "Medium" and "High" consumption classes. The Parent Node of area C splits the data on the basis of U-value of the windows. If it is lower than 4.0 W/m^2K the energy demand is classified as "Medium". Additionally, 21.5 % of the dwellings (3,351) grouped in this Leaf Node were built before 1960 and the average U-value of window of these sample is lower than 2.5 W/m^2K. It can therefore be deduced that the windows were subject to a refurbishment.

4.2. Classification accuracy

According to [11], in a classification process, a minimum confidence of 50 % ensures the reliability of each Leaf Node. In [12] and [7] the accuracy of the whole classification process is considered acceptable where the uncertainty is lower than 20 – 30%. In this case, 83.7 % of the sample was correctly classified, demonstrating a reliable accuracy. The best classified category includes the "High" consumption dwellings, also because of features easily recognizable. On the contrary, the worst classified samples belong to the "Low" consumption class. This result was predictable mainly because of the intrinsic definition of this class. Indeed having to include the best 5 % performing samples, the dimension of this cluster is significantly lower than the others and some affecting variables could be neglected. Nevertheless, 70.6 % of samples in this class are rightly classified and this accuracy is still acceptable. Moreover, in some simulations herewith not presented, the thresholds between two classes were changed in order to test the dependence of the results accuracy by the selection of these values. In any analysed test, the classification process was included in 75-85 % range of accuracy. The final classification tree was selected because it included the most significant threshold statistical values as described in the Section 2.2. Moreover, misclassification between extreme classes ("Low" to "High" and vice versa) were not present . Lastly, remaining 16.3 % of inaccuracy of the model can be amply explained.

Considering an additional cluster which includes the 8,440 "Medium" samples wrongly classified as "High" and 5,131 "High" samples wrongly classified as "Medium", the average nPED of this cluster is 199.24 kWh/m^2 (standard deviation 24.07 kWh/m^2). This value is very close to the selected threshold (205.54 kWh/m^2) which represents the borderline between the "Medium" and "High" dwellings. The same reasoning can be repeated considering the "Low" and "Medium" samples wrongly classified. In this case the cluster of this samples has an average value of 78.60 kWh/m^2 (standard deviation 8.12 kWh/m2), whereas the threshold limit between "Low" and "Medium" was set to 82 kWh/m^2. This misclassification is linked to the high density of the nPED data. Indeed, the classification tree is not able to split two different classes considering a specific value of a dependent variable that can be considered continuous. In proximity of the borders between two consumption categories, two bands of uncertainty were considered. In these bands the nPED average values are very closed to the threshold between different categories, as shown by the small standard deviation. For this reason it is possible to affirm that the misclassified cases belong to an intermediate energy category in which the variables of CART are not able to split the samples again. The number of intermediate bands is dependent on the number of energy categories but it is not dependent by the selection of the thresholds that define the categories.

Furthermore, some of the misclassification drawbacks are due to the restricted number of variables considered in the classification algorithm. Indeed on one hand, the lower the number of variable the simpler the usability of the model. On the other hand, a low number of variables might cause the neglecting of some physical processes. Considering "Low" samples misclassified as "Medium", some of these drawbacks are due to the neglecting of the data regarding ventilation need. In fact, to split the samples the decision tree does not use any variables related to the efficiency of a potential mechanical ventilation heat recovery system installed. Indeed, it is commonly known that for low consumption buildings ventilation represents an important voice to be considered for the evaluation of nPED. In any case, the ventilation rate was not selected as an input of the system, because it mainly affects only a minimum number of samples in the considered dataset.

4.3. Benchmark and Cluster analysis for high consumption dwellings

The dwellings labeled as "High" nPED require a deep attention since in this category high energy saving opportunities exist. In this class are included 45,568 samples. An important step to promote efficient use of energy is to establish an energy performance value (benchmark) and to identify the dwellings in this class that most need energy improvements [10]. A benchmarking analysis was performed considering all the variables influencing the nPED. For all the input attributes (aspect ratio, U-value of opaque envelope and windows and global efficiency of the heating and DHW system), the statistical function such as mean, median, 25th and 75th percentile number were calculated. The intersection of interquartile ranges allowed to select a single reference real dwelling where the attributes are close to the median values. In particular, Table 4 reports the benchmark values for the "High" nPED class.

Table 4. Real Dwelling Reference (Benchmark)

Variables				
S/V [m^{-1}]	Uo [W/m^2K]	Uw [W/m^2K]	η [-]	nPED [kWh/m^2]
0.64	1.20	3.94	0.65	285.51

The benchmark values related to the input variables were used to define thresholds for the identification of clusters which grouped "High" nPED dwellings requiring energy refurbishment. A cluster analysis was performed and the evaluation of Davies-Bouldin index for different configurations showed that the k-means algorithm with 3 clusters produced the best clustering output. Figure 5(a) shows the z scores of benchmark and the vector components of the centroids for the 3 clusters. Clusters 1 and 2 have similar nPED, lower than the cluster 3. The nPED of cluster 3 is mostly due to the highest aspect ratio of the samples. The aspect ratio directly affects the amount of energy losses through the building envelope. Therefore, the improvement of the U-value of opaque envelope and windows for the pattern grouped in this cluster was recommended, whereas dwellings with the worst characteristics of building envelope (transparent and opaque) and system efficiency were grouped in the cluster 2. For this cluster, the low values of nPED was justified only by their low aspect ratio factor (S/V). It can be inferred that to improve the energy efficiency of these dwellings an intervention on the performance of the heating plant system is necessary. Frequency distribution of nPED reported in Figure 5(b) shows the location of clusters compared with the energy benchmark previously calculated. The benchmark of nPED divided the samples into two parts. In general, the nPED of dwellings grouped in cluster 1 and 2 is lower than the benchmark value, while for the dwellings included in cluster 3 is higher. The use of benchmark values on nPED allows to easily select the dwellings which need to be refurbished. However, the causes of high consumptions that are better highlighted through the elaboration shown in Figure 5(a).
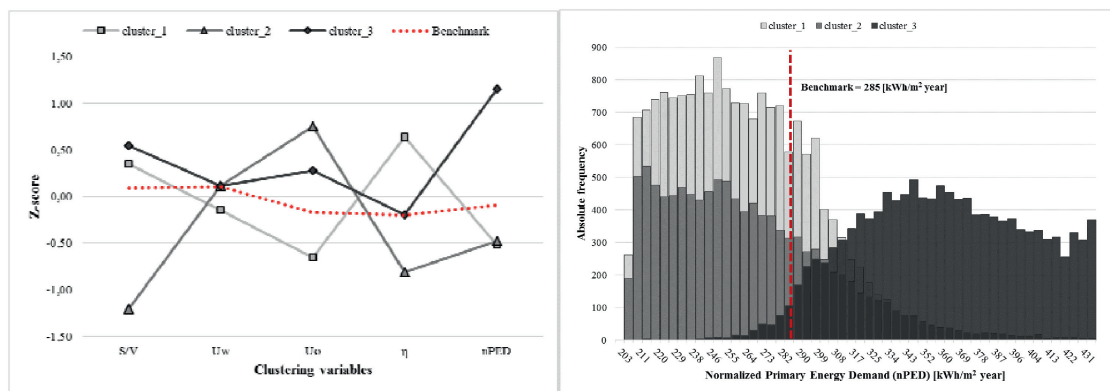

Fig. 5. (a) vector components of cluster centroids; (b) annual demand of PED for high consumption clusters.

5. Conclusions

In this paper, a classification process involving 92,906 dwellings located in Piedmont region was conducted. The influence on the normalised primary energy demand (nPED) of four influencing variables (aspect ratio, U-value of opaque envelope and windows and global efficiency of the heating and DHW system) was analysed. A CART algorithm was developed to classify the dwellings in different categories of nPED ("Low", "Medium" and "High"). Results show that 83.7 % of the training records were correctly classified. Thus, the accuracy of the proposed classification process is very high, despite a restricted number of input variables.

Reference values for the analysed input variables which determine a "Low", "Medium" or "High" nPED were found. The average U-value of the opaque envelope proved to be the most important variable, along with aspect ratio. The method herein presented provides a simple tool for designers and building stakeholders to predict categorical variables related to nPED and to set reference threshold values for physical variables. Due to the large dimensions of the adopted dataset, the information provided can be considered representative of the Italian residential dwelling stock. Moreover, the proposed classification criteria are based on statistical variables easy to be adaptable to different datasets. The CART provided a set of useful decision rules that can determine the pattern that drives the evaluation of the nPED.

Further analysis of the 45,568 most energy-consuming dwellings (classified as "High") was carried out to provide some useful information on the basis of the input variables. Through a frequency distribution analysis for the "High" nPED category a reference dwelling was defined, considering the intersection of the interquartile range among all the variables. This process provided a single real dwelling that represents the benchmark of the "High" energy consumption dwellings. In particular, this real dwelling was useful as a reference for establishing priorities in the selection of dwellings needing energy refurbishment. Furthermore, it is helpful in allowing designers to choose the variables that most influence energy consumption. A cluster analysis was performed in order to establish the intrinsic characteristic of "High" nPED. Furthermore, by applying the k-means algorithm to the same sample, it was possible to evaluate the design variables that need to be taken into account in a refurbishment process for each defined cluster in this category.

The accuracy of the classification method proved reliable, compared to the typical values suggested by the literature. However, aside from some crucial variables (thermal bridging effect, opaque/transparent ratio, mechanical ventilation and MVHR), the model does have limitations. Future works will focus on detailed investigations relating to specific subsets and will also consider some variables that were not taken into account in the present paper, evaluating their influence on PED.

References

[1] Kim H, Stumpf A, Kim W. Analysis of an energy efficient building design through data mining approach. Automation in Construction 2011; 20:37-43.
[2] Xiao F, Fan C. Data mining in building automation system for improving building operational performance. Energ Buildings 2014; 75:109-118.
[3] Khan I, Capozzoli A, Corgnati S P, Cerquitelli. Fault detection analysis of building energy consumption using data mining techniques. Procedia energy 2013; 42:557-566.
[4] Capozzoli A, Lauro F, Khan I. Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Syst ppl 2015; 42:4324-4338.
[5] Ballarini I, Corgnati S P, Corrado V, Talà N. Improving energy modeling of large building stock through the Development of archetype buildings. 12th Conference of the International Building Performance Simulation. Australia, Sydney 14-16 November 2011.
[6] Summerfield A J, Raslan R, Lowe R J, Oreszczyn T. How useful are building energy models for policy? A UK perspective. 12th Conference of the International Building Performance Simulation. Australia, Sydney 14-16 November 2011.
[7] Yu Z, Haghighat F, Fung B, Yoshino H. A decision tree method for building energy demand modeling. Energ Buildings 2010; 42:1637-1646.
[8] Ballarini I, Corgnati S P, Corrado V, Talà N. Definition of building typologies for energy investigations on residential sector by Tabula Iee-Project: Application to Italian case studies. Roomvent, Trondheim 19-22 June 2011.
[9] C. Charu Aggarwal. Data classification algorithm and application. New York: Chapman & Hall/CRC; 2014.
[10] Mikucioniene R, Martinaitis V, Keras E. Evaluation of energy efficiency measures sustainability by decision tree method. Energ Buildings 2014; 76:64-71.
[11] D'Oca S, Honga T. Occupancy schedules learning process through a data mining framework. Energ Buildings 2015; 88:395-408.
[12] Gao Y G Y, Tumwesigye E, Cahill B, Menzel K.. Using data mining in optimisation of building energy consumption and thermal comfort management. 2nd International Conference on Software Engineering and Data Mining. China, Chengdu, 23-25 June 2010.