2014

Selection of Representative Buildings through Preliminary Cluster Analysis

Rigoberto Arambula Lara
Free University of Bozen/Bolzano, Italy, rigoberto.arambula@natec.unibz.it

Francesca Cappelletti
University IUAV of Venezia, Italy, francesca.cappelletti@iuav.it

Piercarlo Romagnoni
University IUAV of Venezia, Italy, pierca@iuav.it

Andrea Gasparella
Free University of Bozen/Bolzano, Italy, andrea.gasparella@unibz.it

Preliminary Cluster Analysis for retrofit actions planning

Rigoberto ARAMBULA[1]*, Francesca CAPPELLETTI[2],
Piercarlo ROMAGNONI[3], Andrea GASPARELLA[4]

[1,3] Free University of Bolzano/Bozen, Faculty of Science and Technology, Bolzano, Italy
+39 0471017619, rigoberto.arambula@natec.unibz.it

[2,4] University Iuav, Department of Design and Planning in Complex Environments, Venice, Italy
francesca.cappelletti@iuav.it

* Corresponding Author

ABSTRACT

The present work refers to a group of 59 schools located in the North Italian province of Treviso, for which metered energy consumption and seasonal degree days were available for the last five year period. Geometrical features of each school, such as the gross and net heated volume, the floor area, the window area, and the dispersing envelope surface were also known. Moreover, data about the thermal resistance of the building envelope components and the type of heating system were available. For each school, energy and geometric indicators have been calculated: the ratio between dispersing area and gross heated volume, the window to wall ratio, the energy consumption per volume unit and the energy per volume unit and Heating Degree Hours (HDH).
To characterize the features and performance of the buildings, and to assess the possibility to select a sample of representative schools to be further monitored, a cluster analysis has been conducted. The main issues to be solved in order to develop this analysis are the definition of the type and the most suitable number of parameters to be correlated to energy consumption, and the determination of the adequate number of clusters.
At first, the available parameters have been grouped in all possible combination sets from 2 to 8 elements and a multiple linear regression was calculated for each single configuration, in order to express the level of dependence of the school total energy consumption on a specific set of parameters. Since the coefficients of determination changes are negligible for more than 6 parameters, this seemed to be an acceptable compromise between representativeness and complexity.
The sets of parameters, which better explain the energy performance, have been determined by considering the best results from the regressions. K-means cluster analysis was then performed on the school sample, considering the parameters in those sets, in order to find 3 clusters according to each parameters' set.
Regression analysis has been repeated for every group, to check if the correlation between parameters and energy consumption improves inside each cluster with respect to the whole sample. The same method was then repeated to find sub-clusters in those groups of buildings with the lowest correlation coefficients in order to divide them into more homogeneous groups, with a higher correlation between the buildings characteristics and their final energy consumption.

1. INTRODUCTION

Educational buildings of the European stock have been built in quite different periods over a very large time span, resulting in a heterogeneous built environment. Italy is no exception, while about 60 % of the 42 000 occupied school buildings were built before 1974, and even if around 50 % of them benefited from major maintenance works during the last decade, as much as 30 % of the building stock still requires retrofit interventions (Legambiente, 2012) to achieve the level of energy performance and indoor comfort conditions established in recent regulation.

Awareness of the need to improve the existing buildings´ comfort conditions and energy performance has been rising in the last three decades. Numerous studies have been carried out to determine both the real dimension of the problem and to propose technically and economically feasible solutions, while governments have established tougher regulations and standards to be complied by new and retrofit construction.
Nowadays the debate concerning the energy retrofit of existing buildings is oriented to the research of the most convenient retrofit actions from an economic point of view. The methodology to be implemented to reach this objective consists in a cost-optimal analysis of different retrofit improvements, starting from a reference building which has to be representative of a building category. Defining the reference building in a stock of existing ones implies the analysis of a large amount of information to find out how this sample can be grouped.
Many studies on building stock classification and benchmarking have been carried out, some of them concerning school buildings in particular.
Starting from available data of 1100 schools in Greece, Gaitani et al (2010) defined the main components to select a group of representative buildings to perform further studies: heated area, age of the building and heating system, envelope insulation, number of classrooms and students and occupancy profile. In order to reduce the number of variables analyzed together, contribution of each to the final energy performance was calculated individually. Using a different approach, geometric configuration was included by Dimoudi and Kostarela (2009) as a variable to select a representative sample of school buildings in Greece, in order to evaluate the effect of different interventions, both individually and combining them.
Several methods, using auditing to assess the consumption of large groups of buildings with the scope of defining a benchmark, can be found. Desideri and Proietti (2002) chose to calculate energy consumption indexes to classify 29 schools in the central-Italy province of Perugia. More specific research was carried out on schools, whose performance was far from the average, in order to determine the causes of the energy consumption differences and to define possible interventions. Moreover, they calculated the savings that could be achieved if the proposed improvements were applied to all school structures in the province. Hernandez et al (2008) proposed a method to calculate the energy performance benchmark for a rating system using a calculated energy performance indicator and grading it according to standard EN 15217:2007. A group of primary schools in Ireland was used as case-study and the main problem they encountered was the lack of historical data, a problem that is also found in Italy.
In some cases the building stock is very large requiring the application of some statistical techniques in order to group buildings with homogeneous characteristics. Many data mining algorithms can be used in order to find correlations and patterns. One of such techniques is clustering analysis, by which a set of elements is split into several homogeneous groups containing elements that are similar to each other and significantly differ from those of any other group. Cluster analysis has been already used as a tool for the classification of large building samples, although considering a single variable for the definition of the groups: Santamouris et al. (2007) used fuzzy clustering techniques to define energy classes based on heating energy consumption of a large sample of schools in Greece. Similarly, Gaitani et al. (2010) classified schools by means of k-means clustering, considering the normalized thermal energy consumption. For the analysis of the Chilean housing market, nine apartment typologies were defined by Encinas and De Herde (2013) using cluster analysis on a database containing thousands of units. Some authors have applied clustering analysis to evaluate different aspects, such as the building typologies in a building stock (Famuyibo et al., 2012) or even the occupant behavior in relation to the energy loads.
When available, buildings energy consumption data have to be correlated with the buildings characteristics, inferring cause-effect relations, to explain the reasons for a given performance and to support the definition of improvement measures. While being a relatively simple issue for a single building, this could be a very difficult and time-consuming task when a large stock is considered. In addition, even when consumption data are available, information about buildings characteristics is often very limited.
The aim of this work is to explore the possibility of supporting the energy audit of a large building stock using a few synthetic descriptors, calculated for homogeneous groups defined by means of clustering. To achieve this objective, a method to determine the buildings characteristics having the highest contribution on final energy consumption levels, is presented. A sample of 59 schools, located in the Province of Treviso, in the North-East of Italy, has been analyzed. Extensive building data has been elaborated, using regression techniques as well as clustering analysis, in order to define the groups of parameters that are better correlated to the final energy consumption. This is the first step of a wider research project that aims to the selection of a few schools that represent the whole sample, in order to perform a detailed energy audit of them and evaluate the real effects of previous retrofit actions, as well as the expected results of new interventions. Monitoring of indoor comfort indicators, modelling and simulation using calibration techniques will be used to optimize the possible energy performance improvement strategies. Although in this case the sample is composed of schools, this methodology is meant to be applied on groups from different building typologies as well.

2. DESCRIPTION OF THE SCHOOLS SAMPLE

2.1 Geometrical and thermal characteristics

A large database containing information from 85 high school buildings owned by the province has been analyzed. For each building, data regarding geometry, thermal properties of the building envelope and energy consumption of 5 years, from 2008 to 2013, were available. After a first control, some of the buildings have not been considered for further analysis, because of missing or inconsistent information. The final selected sample includes 59 buildings.

Buildings, built before the publication of the first energy consumption regulation law in Italy (Law 373/1976), account for about half of the sample, and as much as 75 % of the schools are under 20 000 m^3 of gross heated volume. Almost 90 % of them have natural gas heating systems, and 40 % of them have an installed heating power in the 300-600 kW range.

With respect to the weather conditions, the Province of Treviso is located in the Italian climatic zone E (Cfa according to Köppen classification). Schools are situated in different locations with a conventional number of Heating Degree Days (HDD) spanning from 2350 to 2700.

The frequency distribution of the schools concerning the ratio between the dispersing area and the heated volume (S/V ratio) shows that the sample is composed mostly of quite compact buildings with a S/V ratio varying from 0.3 to 0.5 (Figure 1, left).

Concerning the thermal transmittance of the components, most of the schools (80 %) have non insulated envelopes with average U-value over 0.7 W/(m² K), while almost 42 % of schools have quite good windows with average thermal transmittance under 2.5 W/(m² K). In Figure 1 (right side) the frequency distribution of the average transmittance of building envelopes is plotted: as it can be seen around 60 % of the sample has an average envelope U-value higher than 1 W/(m² K).

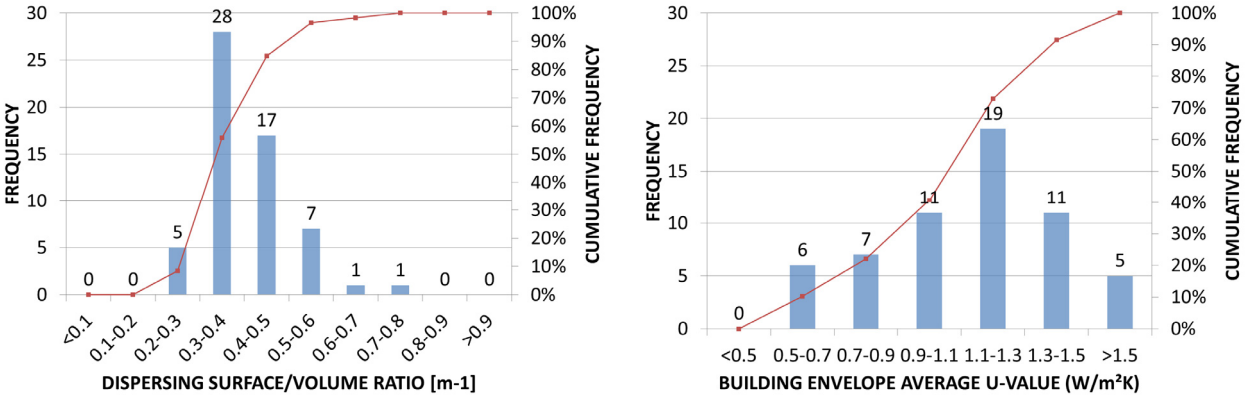

Figure 1: Frequency of buildings for S/V ratio (on the left) and average envelope U-value (on the right).

2.2 Energy consumption

Energy consumption data of the last 5 years (2008-2013), as collected by the school management service, were available for all buildings. Meteorological data for the same period coming from 10 monitoring stations administrated by the regional environmental agency (ARPA Veneto) in different locations through the province were collected. Since the occupancy period is limited to school year 2011-2012, the consumptions of this year were used in the following cluster analysis. In order to compare the energy performance of schools the energy consumption of each of them has been normalized respect to heated volume and heating degree hours. Figure 2 shows the trend of the energy index which is very variable: from 10 Wh/(m^3 Kh) to 150 Wh/(m^3 Kh).

2.3 Occupancy

Occupancy schedules of all the schools were available only for the academic year 2011-2012. Occupancy schedule changes according to the specific school regulation, depending on the type of school and on the extra-school activities that take place in the building outside the teaching timetable. These schedules were used to calculate the total occupancy hours during the corresponding heating period (October 15th-April 15th) for each school.

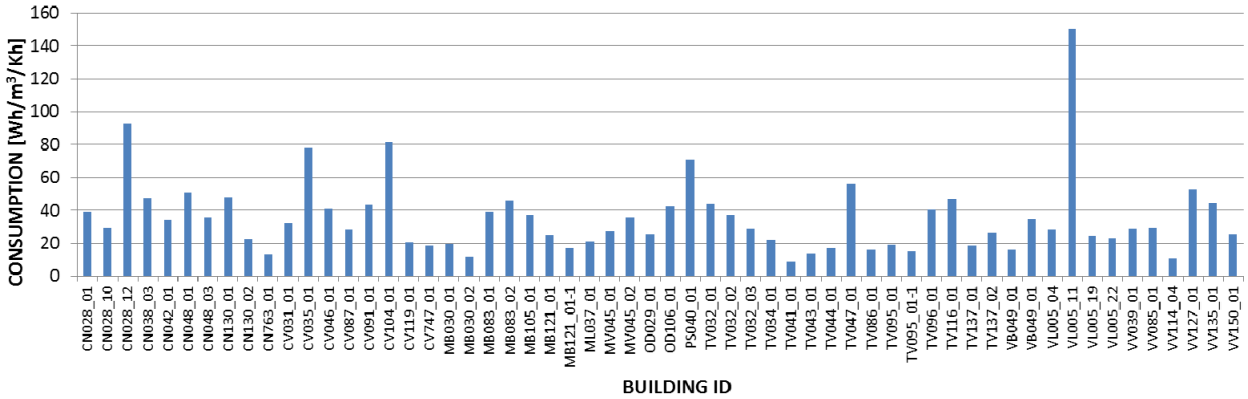

Figure 2: Annual energy consumption per unit heated volume and per unit degree hour (scholastic year 2011-2012)

Since during occupancy hours heating has to be provided in order to maintain the indoor air temperature at the set-point, it was possible to calculate the specific Heating Degree Hours (HDH) for the scholastic year 2011-2012. As shown in Equation 1, HDH were calculated as the sum of all the hourly differences between the external air temperature and a supposed internal set-point temperature of 20 degrees during every occupancy hour of the analyzed heating period.

$$HDH = \Sigma\ (T_{int} - T_{ext}) \cdot h \qquad (1)$$

In Figure 3 the results of this calculation are plotted, showing that the range of values present in this group of buildings is quite ample, as it goes from around 500 to 1400 HDH.

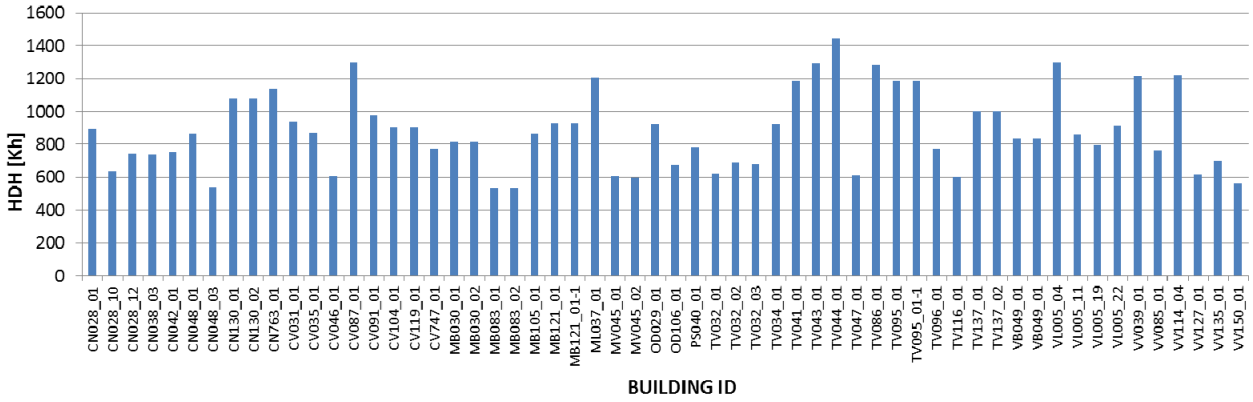

Figure 3: Heating degree hours (HDH) during heating period for the scholastic year 2011-2012

3. FIRST REGRESSION ANALYSIS

Every building has a particular energy consumption demand resulting from the combination of many variables, spanning from location and geometry to occupancy and wall stratification. However, the contribution of a small number of those variables is much higher than the rest. Since the aim of the work was to group schools with similar characteristics and similar correlations between variables and consumption, the first issue was to determine which parameters, both individually or in combination with others, have a stronger correlation with the specific energy consumption normalized on the heated volume and heating degree hours.

Building energy consumption per heating degree hour and per cubic meter of heated volume was defined as the dependent quantity (response) while the 12 independent parameters (predictors) listed in Table 1 were the ones used to compose the different combinations.

To evaluate the correlation of parameters with energy consumption, multiple linear regression was calculated for several combinations of variables. Predictors have been grouped in all the possible combinations containing from 2

to 8 parameters. Afterwards, a multiple linear regression was calculated for each single configuration set, in order to estimate the dependence of the building energy consumption on that particular combination of parameters. Configurations with a larger number of predictors provide higher adjusted coefficients of determination (R^2_{adj}). However, adding one parameter to a set of 6, does not increase significantly the correlation. For this reason combinations of 6 parameters, in addition to the energy consumption, were chosen for clustering. The 10 configuration sets with the highest correlation values between predictors and response have been selected to perform the cluster analysis. Each of those sets of parameters is identified with an identification number, ID. In Table 1 the parameters included in each group are listed; configurations have been ordered with respect to their R^2_{adj} value. This order will be used for the rest of the present paper.

Table 1: Parameters included in the 10 selected configurations with higher R^2_{adj} including all buildings in the sample.

CONFIGURATION ID	235	825	311	53	304	309	307	270	12	242
R^2adj	0,298	0,283	0,279	0,277	0,276	0,276	0,275	0,275	0,272	0,270
EXTERNAL WALL	■		■	■	■	■	■	■	■	■
ROOF AREA				■					■	
GROUND FLOOR AREA	■	■	■	■	■	■	■	■	■	■
OPAQUE ENVELOPE	■								■	■
TRANSPARENT ENVELOPE								■		
AVERAGE U VALUE	■	■	■	■	■	■	■	■	■	■
S/V RATIO	■	■	■	■	■	■	■	■		
WINDOW TO WALL RATIO		■			■					
WINDOW TO FLOOR RATIO							■			■
FLOOR AREA						■				
HEATING POWER	■	■	■	■	■	■	■	■	■	■
TRANSPARENT/OPAQUE		■	■							

It is to be noted that some parameters could be directly taken into account as important for the consumption definition, because they are found in almost all of the configurations with a higher R^2_{adj}: External wall area is included in 9 different configurations, S/V ratio is in 8 out of 10, whereas Ground floor area, Average U value and Heating power are present in all of them.

4. FIRST CLUSTER ANALYSIS

For each selected configuration, a k-means clustering analysis has been conducted in order to classify the buildings according to all of the parameters included in a particular configuration. K-means is a simple partitional clustering algorithm that attempts to find K non-overlapping clusters (Wu, 2012). By this method, K centroids are selected, according to the desired number of clusters and data points are assigned to the closest centroid according to the squared Euclidean distances from the closer centroid. Every point in the data set is described by means of its coordinates. In our case coordinates are the 6 independent variables found by the multiple linear regression.

In this way, the analyzed schools are determined as 6-dimensional points. Consumptions are not considered to run the first cluster analysis because the scope is to find groups of buildings with similar characteristics independently of the energy consumption levels that they actually have.

Optimization techniques were used in order to control some cluster characteristics: as part of the clustering algorithm, the first centroids are randomly positioned and then moved, at every algorithm iteration, until the sum of the measures, between each centroid and the points included in that particular cluster, is minimized. As a result, starting points could influence the way in which some of these clusters are composed. Some constraints are imposed for the calculation of the first centroids coordinates defining an influence area for each centroid, as a percentage of the total size of the data cloud (30% in this case), in order to prevent cases in which first guess-centroids are too far or too close to each other, and also to partially avoid including very distant data points in the same cluster.

Once established the initial conditions, cluster analysis was calculated to divide the original 59 building sample into 2, 3 or 4 different clusters. After analyzing the resulting number of buildings contained in each group, it has been decided to use 3 clusters (Table 2). In the case of 2 clusters, the sample is divided in groups that still contain a

relatively large number of schools, resulting in a small increase of the determination coefficient. On the contrary, when dividing the sample into 4 clusters, some of them are too small and the quantity of buildings is smaller than the number of parameters to be correlated, thus preventing the calculation of the coefficient of determination. Clustering into 3 groups results in medium sized groups in most of the cases, big enough to be evaluated using R^2_{adj}, and small enough to contain quite similar buildings.

5. REGRESSION ANALYSIS OF EACH CLUSTER

To assess the efficacy of clustering for each cluster a multiple linear regression was calculated, using the same dependent value (energy consumption) as in the first regression. The objective of the calculation was to evaluate the improvement (if any) of the correlation when considering a more homogeneous group of buildings. It was found that the determination coefficients have increased in comparison with the former R^2_{adj} values, calculated considering all buildings of the sample. This is because the schools inside a cluster have many similarities, determining an energy consumption profile that characterizes the whole group. In Table 2 clusters are listed for decreasing R^2_{adj}.
As it can be seen, results regarding determination coefficient vary from one cluster to the other, because of the relative similarity (or distance) between their members. While cluster 1 typically shows R^2_{adj} values around 0,9, larger variability is shown for the other ones. In order to select the most adequate configurations to explain the energy consumption, some criteria have been established as follows: a significant value of R^2_{adj} (higher than 0,5) in at least one of the clusters and an adequate number of buildings in the remaining clusters to perform a second clustering analysis. According to these criteria, 3 configurations (ID 304, 307 and 270) were selected to continue the analysis. These configurations are also characterized by a very similar composition in at least two of the three clusters: clusters 1 are exactly the same, while clusters 3 differ in a small number of elements.
To compare the clusters from one configuration to another and to explain the difference between centroids, the parallel coordinates plot was used. This kind of plot (Figure 4) can visualize multivariate data for each building on parallel axes, with a line that represents the normalized values for each parameter. The normalization is obtained by dividing each variable by its maximum value. The coordinates of each centroid are representative for a given group of multivariate data points, for a given period of time, and show the peculiarities of the clusters. Considering for instance the configuration 304, in cluster 1 there are buildings with higher external wall area, ground floor area, Window to Wall Ratio (WWR) and heating power, while having smaller S/V ratio and medium envelope U-value, thus meaning that buildings of this cluster are the biggest of the sample. Cluster 2 and 3 include smaller buildings with some similar features (one to the other) such as the S/V ratio, the ground floor area and the heating power. Cluster 2 has the highest envelope average U-value and the lowest WWR. The distance from each school in selected clusters to its centroid was calculated and schools that were closer have been analyzed. In cluster 1 the school closest to the centroid (TV043-01), for configuration 270, is a quite large building of about 43 000 m^3 of heated volume, a heating boiler of 1608 kW, 30973 K h HDHs and its energy consumption is 0.5 Wh/(m^3 K h). In cluster 2 the school closest to centroid in configurations 304 and 270 (CN048-03) is a quite small building of about 5000 m^3 of heated volume, with a heating boiler of 378 kW, 12971 K h HDHs and 1.5 Wh/(m^3 K h). Cluster 3 is the largest one. Though it has the schools closest to each other, nonetheless it has the worst coefficient of correlation. All the configurations have the same closest school (MB083-02).

Table 2: Multiple regression results from the 3 clusters of the 10 selected configurations, including number of buildings per cluster.

	Cluster 1				Cluster 2				Cluster 3			
ID	**R^2_{adj}**	F	p	buildings	**R^2_{adj}**	F	p	buildings	**R^2_{adj}**	F	p	buildings
235	0,6002	5,3687	0,0025	27	0,2828	1,4808	0,2467	24	n/a			7
825	0,4998	3,2792	0,0268	23	0,1850	1,1056	0,3971	27	n/a			8
311	0,8822	8,4663	0,0100	14	0,1216	1,2594	0,3075	37	n/a			7
53	0,7942	3,6827	0,0870	12	0,2074	2,4438	0,0438	44	n/a			2
304	0,9028	6,9866	0,0404	11	0,7544	5,3115	0,0133	18	0,2249	1,5365	0,2151	29
309	0,7782	0,5848	0,7612	9	0,4026	2,1676	0,1051	22	0,2620	1,6196	0,1961	27
307	0,9687	11,5610	0,0817	11	0,4465	2,5382	0,0671	22	0,2462	1,3591	0,2862	25
270	0,8950	4,8116	0,1126	11	0,7691	6,3841	0,0055	17	0,2895	1,9944	0,1121	30
12	0,9056	7,2139	0,0383	13	0,2478	2,6557	0,0326	41	n/a			4
242	0,5427	4,3178	0,0072	25	0,2957	1,7508	0,1665	26	n/a			7

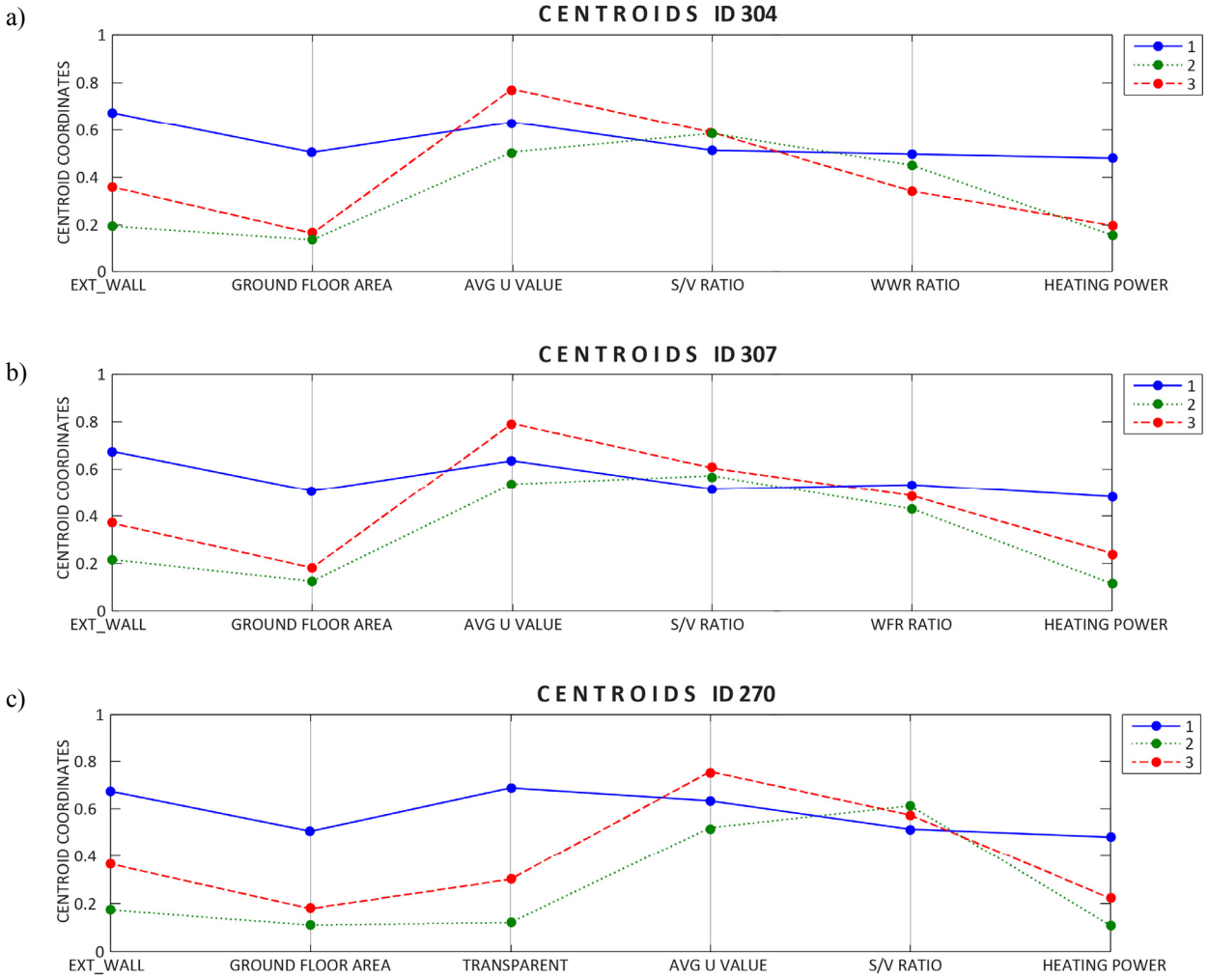

Figure 4: Centroid coordinates (normalized parameters) for first clustering analysis for configuration ID 304 (a), 307 (b) and 270 (c). The normalization is obtained by dividing each variable by its maximum value.

6. SECOND CLUSTER ANALYSIS AND REGRESSION OF SUB-CLUSTERS

In order to improve the correlation of energy consumptions with the independent variables, the clusters with a R^2_{adj} lower than 0.5 (cluster 2 for configuration ID 307 and cluster 3 for all), were divided into 2 sub-clusters each, using the k-means algorithm with the same set-up as for the first clustering, but considering the consumption per volume per degree hour in addition to the 6 variables of the first phase. In this way, according to the results from the first cluster-regression cycle, an increase in the correlation of the sub-clusters was expected. The coordinates of sub-clusters 2.1, 2.2 (cluster 2) for ID 307, and 3.1 and 3.2 (cluster 3) for ID 304, 307 and 270 are shown in figure 5. Looking again to ID 304, sub-clusters 3.1 and 3.2 have new centroids that are similar in average U-value, S/V ratio, WWR, heating power, but differ in external wall area, ground floor area, consumption per volume unit and per HDH. Sub-clusters for cluster 3 of ID 307 and ID 270 show also clear difference regarding the energy consumption levels, heating power and external wall area, while the rest of the parameters are quite similar for each group.
Once the sub-clusters in each selected cluster for a particular configuration (ID) were defined, a new multiple linear regression was calculated to obtain the determination coefficients related to each sub-cluster.

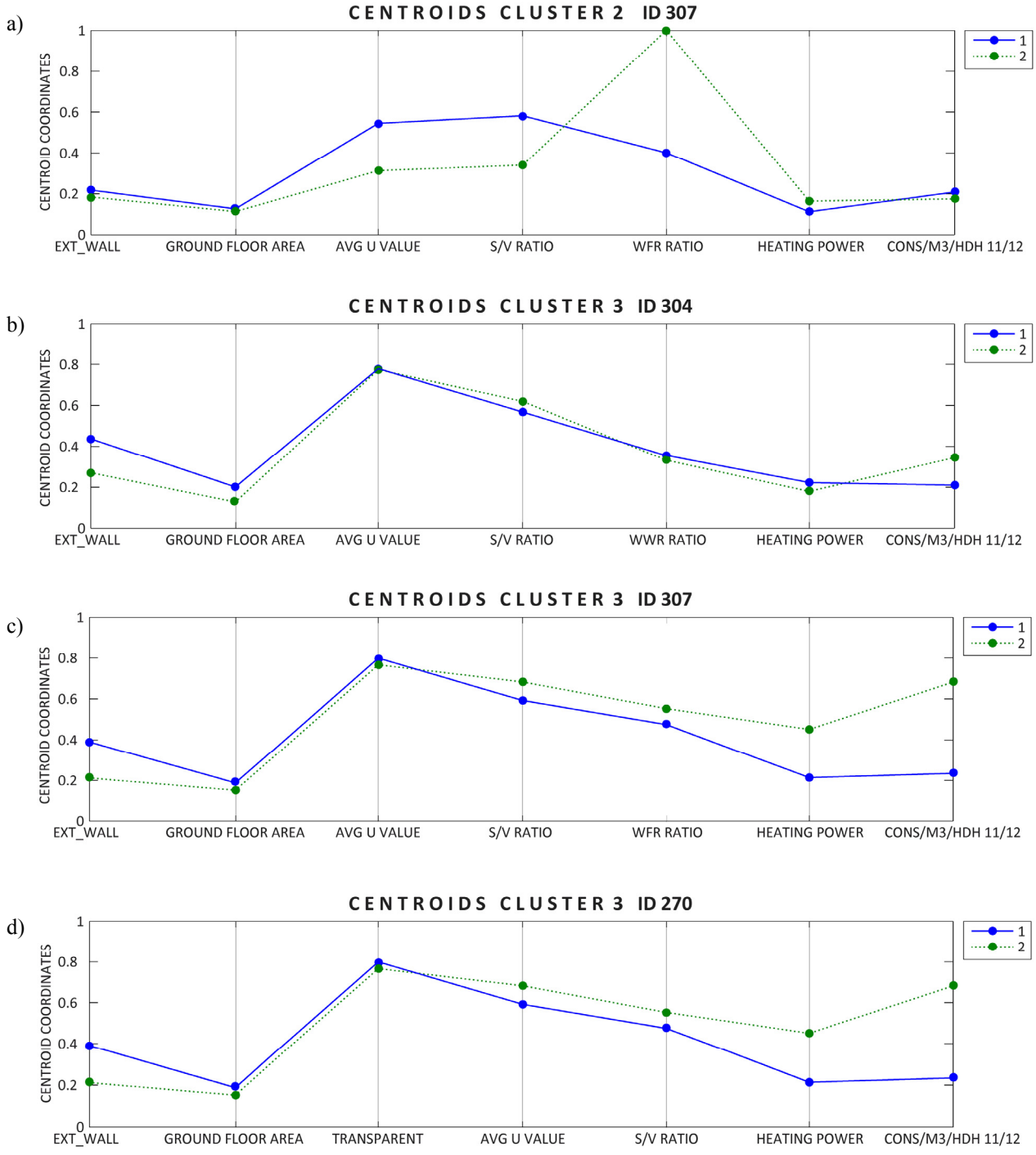

Figure 5: Centroid coordinates (normalized parameters) of sub-clusters for configurations ID 307-2 (a) (2.1 and 2.2), ID 304-3 (b) (3.1 and 3.2), ID 307-3 (c) (3.1 and 3.2) and ID 270-3 (d) (3.1 and 3.2). The normalization is obtained by dividing each variable by its maximum value.

In Table 3 results from the sub-cluster regressions are shown for the combinations ID 304, 307 and 270. Due to the small number of buildings contained in some of the clusters that were divided, in some cases it was not possible to evaluate the correlation using R^2_{adj}. However, for all the reported values (with the exception of sub-cluster 3.1 ID 270), the coefficient of determination further increased. Clusters containing too few buildings, as in sub-cluster 2.2 for ID 307 (1 building) allow to identify buildings that, because of their singular characteristics regarding the considered parameters, should not be analyzed as a group, but using as a case-to-case approach.
Size of each selected cluster and sub-cluster has been determined measuring all building to centroid distances inside every group. Cluster dispersion or density shows how different (or similar) are the buildings included. For instance, cluster 1 of ID 304, which has a high R^2_{adj} (0,9028) contains buildings with distances to centroid between 0,2 and 0,5, with the exception of one building (distance=0,8). Configuration of cluster 2 (R^2_{adj} =0,7544) is similar to the one of cluster 1, having a range of building to centroid distances going from 0,15 to 0,5 with just one school being at 0,9, while all schools in cluster 3 (R^2_{adj} =0,2249) are closer to their centroid, with distances spanning from 0,1 to 0,4, thus making this the most compact cluster. Sub-clusters 3.1 and 3.2 for this ID have a slightly different size, with values from 0,15 to 0,85 for the former and a 0,15-0,95 for the latter.
In the case of ID 307, the 3 clusters are composed similarly, with a main group of buildings very close to their centroid (distances between 0,1 and 0,5), and one school being far away (0,7-0,8) from the rest.
Clusters 1 and 3 for ID 270 are also configured in this way, having distances between 0,2 and 0,5 in cluster 1 and from 0,1 to 0,45 in cluster 3, while cluster 2 is quite compact (0,1-0,4 range).
ID 304 was selected as the most adequate configuration to explain energy consumption within groups due to the high coefficient of determination (>0,5) and low p-values (< 0,05), found at both the first and second clustering levels. For this reason, ID 304 was selected for the continuation of the experimental survey, to be conducted in one representative school from each group with R^2_{adj}>0,5, selected by its proximity to the corresponding cluster centroid.
It is to be noted that p values resulting from the last regression were over 0,05, meaning that for this specific configuration and group of buildings, the resulting coefficient of determination has a low significance. As a result of this, it has been determined as a future development of this method that a different kind of regression analysis has to be used to evaluate this correlation.
Nevertheless, it has been found that the buildings included in the sub-clusters with a small number of schools for configurations ID 307 (sub-cluster 2.2) and ID 270 (sub-cluster 3.2) are similar: 2 schools are located in the town of Castelfranco Veneto and were both built around 1960 (province school codes CV046-01 and CV-119-01), while 1 school is located in Villorba (VL005-19), and has one of the lowest occupancy levels. Regarding ID 307 sub-cluster 1.2, the only included building was built in 1998, being among the most recent from the sample, and it is located in the town of Vittorio Veneto (VV-150-01), the northernmost in the province.
Individual analysis of these schools will be conducted to determine which are the particular characteristics explaining their consumption levels.

Table 3: Multiple regression results from the 2 sub-clusters inside selected clusters

	SUBCLUSTER 2.1				**SUBCLUSTER 2.2**			
ID	**R^2_{adj}**	F	p	buildings	**R^2_{adj}**	F	p	buildings
307	0,4942	2,8126	0,051985	21	n/a			1
	SUBCLUSTER 3.1				**SUBCLUSTER 3.2**			
ID	**R^2_{adj}**	F	p	buildings	**R^2_{adj}**	F	p	buildings
304	0,3611	1,1188	0,42208	16	0,5337	1,392	0,3491	13
307	0,3536	1,8142	0,16354	22	n/a			3
270	0,2945	1,9364	0,12408	27	n/a			3

7. CONCLUSIONS

In this paper a new method for the energy performance classification of existing buildings is presented. The cluster analysis paired with regression techniques has some potential in grouping buildings with similar characteristics. Regression results provide clear parameter comparison and selection elements, as well as useful feedback to the clustering analysis process.
Summarizing, the described methodology has been proved to be suitable for:

- defining which are the parameters and combinations of parameters with a bigger contribution to the final energy consumption of each building in the sample;
- identifying the most suitable parameters to classify a large sample of existing buildings with respect to their energy consumption profile;
- selecting a few representative buildings to be investigated more deeply and for which to individuate the optimal retrofit interventions (for instance through calibrated simulation and multi-objective optimization);

This procedure is useful to find the parameters which better predict the energy consumptions of a large sample of buildings. Buildings with similar parameters can be grouped and the same retrofit strategies should improve, at the same extent, the energy performance of each, thus focusing the interventions on the most relevant aspects and optimizing resources.
Even though the selected type of regression method was not totally suitable to achieve the objective of the work, the development of clustering methodology seems to be promising in the energy classification of a large building stock.

REFERENCES

Antonini E., Boscolo M., Cappelletti F., Romagnoni P., 2009, Schools refurbishment: results of an energy monitoring campaign, *4° Energy Forum*, Bressanone, pp.139-143.

Corgnati S. P., Corrado V., Filippi M., 2008, A method for heating consumption assessment in existing buildings: a field survey concerning 120 Italian schools, *Energy and Buildings*, vol. 40, Issue 5, pp. 801-809

Corgnati S. P., Bellone T., Ariaudo F., 2009, Previsione dei consumi per il riscaldamento ambientale degli edifici esistenti con approccio statistico: il caso delle scuole. *Determination of energy consumption for space heating in existing buildings with statistical approach: the case of schools*, *Proceedings of 64° National Congress ATI*, Italy.

Desideri U., Proietti S., 2002, Analysis of energy consumption in the high schools of a province in central Italy, *Energy and Buildings,* vol. 34, pp. 1003-1016

Dimoudi A., Kostarela P., 2009, Energy monitoring and conservation potential in school buildings in the C´ climatic zone of Greece, *Renewable Energy* , vol. 34, pp. 289-296

Encinas F., De Herde A., 2011, Definition of occupant behavior patterns with respect to ventilation for apartments from real estate market in Santiago de Chile, *Sustainable Cities and Society* 1 pp.38-44

European Commission, *Commission Delegated Regulation EU 244/2012*, *Official Journal of European Union*, L81/18, 20/03/2012, 2012.

Famuyibo A.A., Duffy A., Strachan P., 2012, Developing archetypes for domestic dwellings – an Irish case study. *Energy and Buildings* vol. 50 pp.150-157

Filippin C., 2000, Benchmarking the energy efficiency and greenhouse gases emissions of school buildings in central Argentina, *Building and Environment*, vol. 35, p. 407-414.

Gaitani N., Lehmann C., Santamouris M., Mihalakakou G. and Patargias P., 2010, Using principal component and cluster analysis in the heating evaluation of the school building sector, *Applied Energy*, vol. 87, Issue 6, p.2079-2086.

Hernandez P., Burke K., Lewis J. O., 2008, Development of energy performance benchmarks and building energy ratings for non-domestic buildings: An example for Irish primary schools. *Energy and Buildings*, vol. 40, pp. 249-254.

Legambiente, 2012, Ecosistema Scuola 2012 – XIII Rapporto di Legambiente sulla qualità dell'edilizia scolastica, delle strutture e dei servizi. *XIII Legambiente report on the quality, facilities and services of school buildings.*

Petcharat S., Chungpaibulpatana S., Rakkwamsuk P., 2012, Assessment of potential energy saving using cluster analysis: A case study of lighting systems in buildings, *Energy and Buildings* vol. 52, pp. 145–152

Wu Junjie, 2012, Advances in K-means clustering. A data mining thinking, Springer Thesis, Springer www.springer.com

Yu Z., Haghighat F., Fung B. C. M., Morofsky E., Yoshino H., 2011, A methodology for identifying and improving occupant behavior in residential buildings, *Energy* 36 pp. 6596-6608

ACKNOWLEDGEMENT

We would like to thank the Province of Treviso for making the schools database available for this research.