From Energy Signature to Cluster Analysis: Comparison Between Different Clustering Algorithms

Lorenza Pistore – Free University of Bozen-Bolzano – lorenza.pistore@natec.unibz.it
Giovanni Pernigotto – Free University of Bozen-Bolzano – giovanni.pernigotto@unibz.it
Francesca Cappelletti – University IUAV of Venezia – francesca.cappelletti@iuav.it
Piercarlo Romagnoni – University IUAV of Venezia – piercarlo.romagnoni@iuav.it
Andrea Gasparella - Free University of Bozen-Bolzano – andrea.gasparella@unibz.it

Abstract

The energy audit on the existing buildings has become a priority in the last years, as a consequence of the adoption of the European Directives on building energy efficiency. In particular in Italy, public buildings are often the most inefficient among the stock and, thus, those with the highest potential for improvements. Many methods can be applied to perform an energy diagnosis; one of them is "Energy Signature" simplified method, ES, described in the Annex B of the technical standard EN 15603:2008. The ES can actually be seen as a very simplified model of the building, based on a linear regression between energy consumption and degree-days in a set of reference periods. If applied year after year, the ES allows a fast detection of system faults, changes of use patterns, and to assess the efficacy of different energy management strategies or retrofitting interventions, discounting the effect of weather variations. When the stock of buildings is large, individual energy audits can be too onerous and time consuming and building simulation impracticable. For this reason, ES can be combined with clustering techniques in order to identify groups of buildings with similar behaviour among which a reference case can be identified and deeply investigated either experimentally or through detailed building energy simulation (BES). In this respect, ES and clustering can be seen as the key element to allow the extension of BES also to the analysis of building stocks. In this work, ES and different clustering techniques have been used to analyse a set of 41 schools in the province of Treviso, north of Italy, pointing out the buildings features that most affect their energy signatures through multiple linear regressions. A comparison between two non-hierarchical clustering algorithms, K-means and K-medoids, has been conducted. Particular attention has been paid to the approaches for the evaluation of closeness of schools in the same group and the identification of the reference school for each set. As the final outcome of this research, the impact of the clustering algorithms is discussed, in order to assess to which extent the selection of the schools with the most representative energy signatures can be affected by the choice of the data mining techniques.

1. Introduction

The energy audit of existing buildings is a crucial point in order to identify the potential improvements and interventions to achieve a better energy performance. In particular, the last European Directive (European Parliament and Council of the European Union, 2012) indicated public buildings as pioneers of new energy efficiency policies, because they are both community exposed and, in many cases, the least efficient of the building asset. Among them, school buildings represent an important target: in fact, they have not only to be efficient from an energetic point of view, but also to ensure adequate indoor environmental thermal conditions and pupils' thermal comfort.

In the recent years, building simulation models have been largely used in order to evaluate the energy consumptions of school buildings and the impact of different energy renovation interventions, both from an energetic and an economic point of view. In the research by Niemelä et al. (2016), simulation-based optimization analyses have been implemented to determine cost-optimal refurbishment solutions in typical educational buildings built between 1960 and 1970 in cold climate regions. Stavrakakis et al. (2016) have used dynamic simulation to assess the effect of a cool-roof installation on the

Part of
Pernigotto, G., Patuzzi, F., Prada, A., Corrado, V., & Gasparella, A. (Eds.). 2018. Building simulation applications BSA 2017. bu,press.
https://doi.org/10.13124/9788860461360

thermal and energy performance of a school building in Greece Moreover, model calibration has been implemented using simulated indoor air temperature and the measured one, considering ex-ante and ex-post condition. In the same country, Androulakis et al. (2016) have pointed out the contribution of simulation tools for designing complex heating systems combined with renewable sources. However, when a large stock of buildings is taken into account, a case-by-case energy analysis and building simulation modelling is onerous, time-consuming, and expensive. Consequently, if BES has to be adopted, there is the need for a method to detect buildings' energy behaviour quickly, and to identify the best improvement actions, focusing on a subset of representative buildings. In a previous work (Pistore et al., 2016), the authors proposed a method based on the Energy Signature approach (CEN, 2008) coupled with cluster and multiple regression analyses applied to a stock of 41 schools located in the province of Treviso, Italy. The Energy Signature simplified method is described as a useful diagnostic technique consisting in plotting for several times the energy consumptions for space heating or cooling of a building versus the external temperature averaged over a suitable period, and determining their linear relationship (CEN, 2008). Subsequently, K-means clusterization was applied to group schools in clusters using ES parameters, e.g., the slope of linear regression function and the zero of the function, as dependent variables to group schools in clusters. Homogeneous subsets of schools were identified by means of the buildings' features that most affect the signature parameters, and a building reference case was pointed out for each cluster in order to find the parameters that most affect the energy behaviours and, consequently, the best set of interventions able to improve the buildings' energy efficiency. These efficiency measures can be applied on the reference cases, on which more detailed analyses, such as dynamic simulation modelling, can be performed, and then the solutions can be extended to the other buildings in the same cluster by means of the ES method.

In this framework, the methodology applied for the building stock clusterization assumes a great importance in the success of final goal. In fact, the cluster analysis can be implemented according to different algorithms, which have to be carefully selected without any preconception (Kaufman and Rousseeuw, 2005) since they can be suitable for different purposes according to the type of available data. In some cases, several algorithms are applicable and a priori arguments may not suffice to narrow down the choice to a single method, which leads to the necessity of a comparison between results (Kaufman and Rousseeuw, 2005).

In this paper, a comparison between two widespread non-hierarchical clustering algorithms, K-means and K-medoids (Reynolds et al., 2004) is presented. The comparison is conducted on the same stock of schools analysed by Pistore et al. (2016). Assuming that there is a clear resemblance in the running mode and in the object function of the two clustering algorithms, according to the literature, K-medoids algorithm is considered to be more robust with respect to outliers (Arora and Varshney, 2016; Kaufman and Rousseeuw, 2005; Park et al., 2009). Differently from K-means, the clusters determined following K-medoids are ball-shaped and the representative object (i.e. the medoid) is one element of the dataset. Moreover, various non-Euclidean approaches are available for the calculation of the distances. In the considered case, with a dataset a number of elements n lower than 100, the PAM (Partitioning Around Medoid) algorithm has been adopted, since it is often recommended when the aim is to look for representative objects and characteristics.

2. Method

2.1 Energy Signatures

A building energy audit can be performed by the ES method described in the annex B of the EN 15603:2008 (CEN, 2008). This approach consists in plotting for several time periods the average heating or cooling uses versus the average external temperature, and this allows the user to fast gather useful information about the building energy behaviour and, in a subsequent phase, to verify the refurbishment interventions effect and to forecast the energy consumptions in the further years. In this method, the indoor air temperature is assumed to be constant

and equal to the setpoint (generally 20 °C) during the occupancy time, so that the external air temperature is the most influential parameter. Its application requires gathering data about energy uses for heating or cooling, as well as average external temperatures or, when possible, cumulated differences between actual indoor and outdoor temperatures recorded at regular intervals, from one hour to a week. This latter time period has been adopted in this work, since it is long enough to capture a characteristic occupation or use pattern, while being short enough not to hide climatic variations along the seasons. Weekly natural gas consumptions have been recorded and the mean consumption for hot water production during the non-heating period has been subtracted in order to isolate the energy uses for space heating. The weekly average power per unit of heated air volume, ϕ, obtained by dividing the energy use during one week EP_h per unit of volume V by the number of opening hours per week τ, as in Eq. (1), has been plotted versus the weekly-average temperature differences during the opening hours, $\Delta T_{20,occ}$ as in Eq. (2)).

$$\Phi = \frac{\sum_{i=1}^{7} EP_{h_i}/V}{\tau} \quad [W / m^3] \quad (1)$$

$$\Delta T_{20,occ} = \sum_{i=1}^{n} (20 - T_{ext})_i \quad [K] \quad (2)$$

with n = opening hours of a week

Energy signatures can be characterized by two main parameters: the slope of the regression function, which represents the energy performance of the building, and the intersection with the x-axis, hereafter called zero of the function, which is the minimum temperature difference for which the system is turned on, or the maximum that does not require heating.

2.2 Multiple Linear Regression

According to the methodology already developed and described in Arambula et al. (2015), before implementing clusterization, it is necessary to define a list of independent variables describing the building to be used as predictors of the parameters of the energy signatures. The list of 12 candidate descriptive quantities includes:

(A) the area of the external vertical walls,
(B) roof area,
(C) floor area,
(D) ground floor area,
(E) total area of opaque envelope,
(F) total area of transparent envelope,
(G) windows to vertical walls ratio,
(H) windows to floor ratio,
(I) opaque and transparent envelope area ratio,
(J) average thermal transmittance of the envelope,
(K) envelope compactness ratio,
(L) heating system capacity.

Each quantity is indicated by a letter (A to L) in the next tables. The highest value of each of the 12 descriptive quantities in the building set has been used to normalize the values for each building.

A multiple linear regression has been applied to find the sets of the candidate quantities which better define homogenous groups and to develop the clustering. Starting from groups of 2 to groups of 12 variables, 4083 possible combinations of the 12 normalized descriptive quantities have been defined and used as predictors in multiple linear regression models. For each regression, the adjusted index of determination R^2_{adj} has been calculated, as well as F-tests and the p-values, to check the model's statistical significance, and variance inflation factors VIF, for the analysis of multi-collinearity issues. Only models with significant p-value with respect to a significance level of 10 % and without multi-collinearity issues (i.e. VIF < 10) have been considered for the definition of the quantities for the clustering. The combinations of predictors with the highest R^2_{adj} have been selected as sets of coordinates of each element in the sample of schools. After a preliminary study, the zero of the function has been found poorly correlated to the set of proposed variables and has been discarded from the analysis, focused only on the slope of the energy signatures.

2.3 Clustering and Maximization of the Explained Variance

A comparison between two partitioning methods, K-means and K-medoids PAM has been performed. Both approaches imply that each cluster contains at least one element; each element belongs to only one

cluster and the number of clusters is $k \leq n$, where n is the number of elements to be grouped. Once defined the desired number of clusters k, an equivalent number of centroids or medoids is randomly selected and data points are assigned (Junjie, 2012). In this work, since the whole dataset for the clustering includes 41 elements, k has been imposed equal to 2, to facilitate the definition of groups with $n \geq 12$ in accordance with the statistical central limit theorem. After the attribution of the elements to the different clusters, it is checked if the variance explained by the predictors can be increased further: if for a given cluster there is a combination of predictors with higher R^2_{adj}, statistically significant F-value and p-value and limited VIF, then it is adopted as new coordinate systems for the elements belonging to it. If for a cluster the explained variance cannot be improved but at least 25 elements are present, it is possible to sub-cluster with k equal to 2 using the best set of coordinates available for that cluster. As a final step, the school closest to the centroid (K-means method) and the medoid schools (in K-medoids method) in each cluster have been selected.

The main difference between the two methods are the followings: in the K-means, initial virtual centroids are randomly generated within the domain of the dataset and objects are assigned to the clusters using the square Euclidean distance calculation, which is implemented by the algorithm itself at each iteration. Each cluster is defined around a centretype (the centroid), whose coordinates are the mean of the coordinates of all the cluster's elements. In K-medoids, the calculation of the distances is not repeated in each iteration, but the algorithm seeks distance information from a distance matrix. The representative object of each group, i.e. the medoid, is chosen at each iteration in order to minimize the distances between it and the other elements in the data set, so as to refine the clusters themselves each time.

2.4 Comparison Method

In order to make a comparison between the two approaches, some criteria and indicators have been identified and used:

1. Number of elements in each cluster.
2. Composition of clusters in terms of specific schools grouped in the same cluster.
3. Equivalence of the identified reference buildings in the two methods.
4. Variance explained by the selected variables in the multiple linear regression. For this purpose, the indicator used is the adjusted index of determination.
5. Homogeneity and level of similarities of cluster elements. Two indicators have been considered in this case: the standard deviation from the centroid/medoid for each variable of the elements within each cluster and the sum of the square Euclidean distances in each cluster.

3. Results

The characteristics of the clusters obtained by K-means and K-medoids approaches are reported respectively in Tables 1 and 2. All the top-ten combinations of descriptive quantities (identified through an ID-string, composed by the letters representing the descriptive quantities included) have been used for clustering. Then, one of the combinations has been chosen considering its statistical significance. In particular, R^2_{adj}, F-value, p-value, and VIF have been analysed in order to obtain at least one cluster with significant values, and the other one, even if inconsequential, with such a number of elements that allows a subclustering. For these reasons, configuration AHIJ has been chosen as the initial model for both algorithms as the best performing one, leading to 30 and 11 elements in K-means, and vice versa for K-medoids. From now on, the clusters' names have been assigned with respect to the decreasing order of R^2_{adj} values and number of elements: for example, the cluster with the highest R^2_{adj} and the largest number of elements is CL1.

For K-means algorithm (Table 1), configurations BEGHLI and BCJ are found maximizing the explained variance in the two clusters, with R^2_{adj} respectively equal to 0.591 and 0.679. While the latter can be considered satisfactory and identified as CL1, the former cluster can be divided into 2 subclusters and the explained variance further optimized, using configuration FJ for CL2 and ADGHK

for CL3. CL2 and CL3 are composed of respectively 20 and 11 buildings. Three final clusters have been obtained: CL1 with a R^2_{adj} of 0.679, and CL2 and CL3 with R^2_{adj} respectively of 0.312 and 0.868.

The same approach has been implemented with K-medoids algorithm (Table 2). Configurations CGHIL and BJCLI, both made of 5 variables, maximize the explained variance in the two clusters with R^2_{adj} equal to 0.954 (CL1) and 0.132 (CL2), respectively. The latter cluster has been divided again into 2 subclusters and the explained variance further optimized, using EHKL variables for CL2 and BEHJK variables for CL3. CL2 and CL3 are composed of respectively 10 and 20 buildings, with a R^2_{adj} of 0.939 and 0.311 respectively.

K-medoids algorithm gives higher R^2_{adj} for two over three clusters, while R^2_{adj} of CL2-kmeans and CL3-kmedoids is the same for the two algorithms. Looking at the number of elements in the three clusters, we can highlight that for both approaches we have obtained three clusters of 11, 10, and 20 objects but the distribution of the schools among the groups is different and the two methods are different.

Table 3 reports the standardized coefficients of the building variables included in the linear models defined at each step. In both algorithms, the variables selected by regression for the initial model are changed when the explained variance is maximized for the single clusters. Indeed, the variables included in the final models have the highest explanatory power for each cluster and differentiate each group of schools from the others. Furthermore, these final models could be used, in a next phase, for the preliminary assessment of the energy efficiency measures. Comparing K-means and K-medoids algorithms, it can be observed that both final models and included variables are different.

In Fig. 1 the school position inside clusters is shown with respect to K-means algorithm first, and K-medoids second. As it can be observed, elements generally change from one cluster to another by changing the clustering algorithm and, moreover, also the identified reference buildings change between the two different approaches.

However, in order to identify the best performing algorithm, it is necessary to analyse also the output of the predictive models. A comparison between the standard deviation of each variable with respect to the centroid / medoid in each cluster, and the sum of the square Euclidean distances, have been performed as shown in Fig.s 2 and 3. As it can be observed, in K-medoids clusters, standard deviation and the sum of square Euclidean distances is lower, pointing out a more compactness and homogeneity of the groups of schools.

Table 1 - K-means. Results of the clustering and maximization of the explained variance. In bold those p-values significant with respect to a significance value of 10 %. The red square highlights the final clusters obtained and the best predictive models inside them

	First clustering		Subclustering	
	Initial Model	Best Model	Initial Model	Best Model
ID-string	AHIJ	BEGHLI	BEGHLI-a	FJ (CL2)
R^2_{adj}	0.539	0.591	0.046	0.312
F value	9.773 →	8.213 →	1.154 →	5.299
p-value	< 0.001	< 0.001	0.387	0.016
N	31	31	20	20
ID-string			BEGHLI-b	ADGHK (CL3)
R^2_{adj}			0.896	0.868
F value			15.390 →	14.167
p-value			0.010	0.006
N			11	11
ID-string	AHIJ	BCJ (CL1)		
R^2_{adj}	0.007	0.679		
F value	1.016 →	7.332		
p-value	0.479	0.020		
N	10	10		

Table 2 - K-medoids. Results of the clustering and maximization of the explained variance. In bold those p-values significant with respect to a significance value of 10 %

	Clustering		Subclustering	
	Initial Model	Best Model	Initial Model	Best Model
ID-string	AHIJ	CGHIL (CL1)		
R^2_{adj}	0.632	0.954		
F value	5.299	42.410		
p-value	0.036	< 0.001		
N	11	11		
ID-string	AHIJ	BJCLI	BJCLI-a	EHKL (CL2)
R^2_{adj}	0.038	0.132	-0.323	0.939
F value	1.284	1.882	0.561	35.863
p-value	0.303	0.135	0.731	0.001
N	30	30	10	10
ID-string			BJCLI-b	BEHJK (CL3)
R^2_{adj}			0.191	0.311
F value			1.900	2.712
p-value			0.158	0.065
N			20	20

Table 3 – K-means and K-medoids: involved variables and standardized coefficients of the linear models

		K-means			K-medoids		
ID-string	I regr. AHIJ Coeff.	CL1 ADGHK Coeff.	CL2 FJ Coeff.	CL3 BCJ Coeff.	CL1 CGHIL Coeff.	CL2 EHKL Coeff.	CL3 BEHJK Coeff.
Descriptors							
A	-0.06	-0.59					
B				-2.18			1.55
C				1.98	-1.08		
D		0.71					
E						-1.31	-0.82
F			0.35				
G		0.71			0.89		
H	0.33	1.01			0.14	-0.75	-0.57
I	-0.34				-0.79		
J	0.35		0.69	0.40			0.31
K		-0.45				0.61	0.40
L					0.24	1.07	
R^2_{adj}	0.21	0.87	0.31	0.68	0.95	0.94	0.31
F value	3.64	14.17	5.30	6.87	42.41	35.86	2.71
p-value	0.01	0.01	0.02	0.03	<0.01	<0.01	0.06

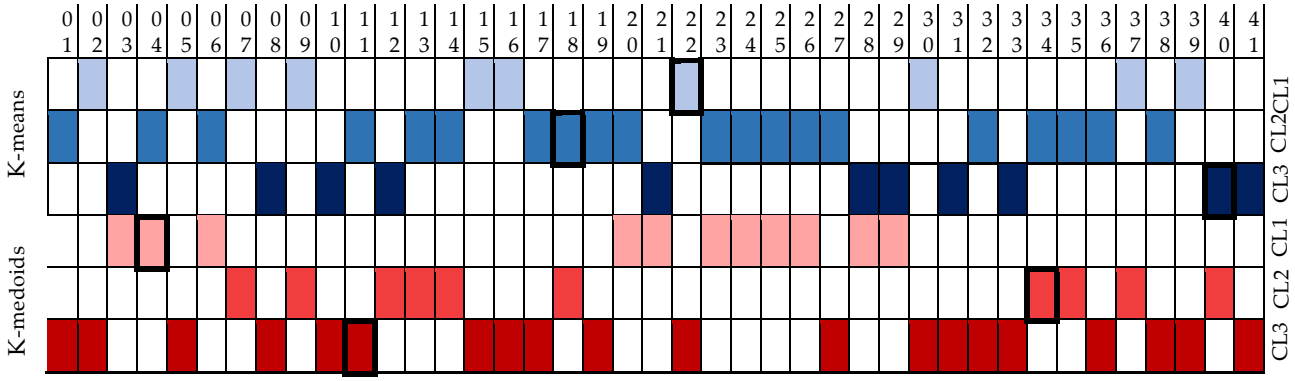

Fig. 1 - Schools position inside clusters by K-means and K-medoids algorithm. The reference buildings are indicated with the thicker black border

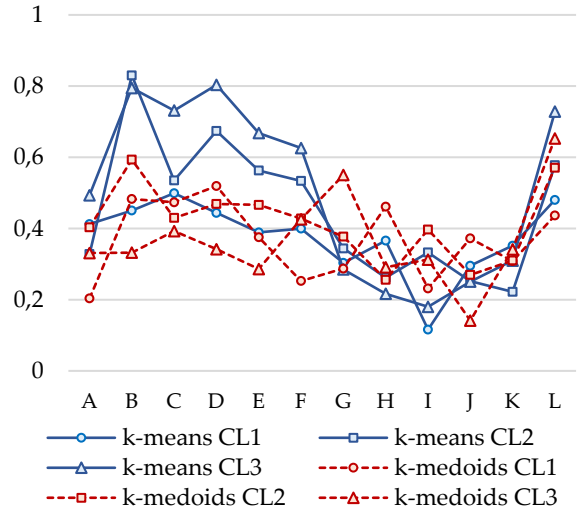

Fig. 2 - Comparison between the standard deviation for each variable with respect to the centroid / medoid of each cluster. K-means in blue colour, K-medoids in dotted red

Fig. 3 - Comparison between the addition of the square Euclidean distances in the three clusters. K-means in blue colour, K-medoids in dotted red

Finally, Fig. 4 compares the energy signature slope predicted by the linear regression models with the largest explanatory power with the actual ones, for each cluster. As it can be observed, two of three models fit better for K-medoids clusters, since a greater number of elements in the stock are within the error band of ±20 %. There is one exception for CL3, which already showed a lower R^2_{adj}.

4. Conclusions

In this paper, a comparison between two different clustering algorithms has been performed. Even if the two approaches run similarly, K-medoids results to be better performing instead of K-means, for the following reasons:

- The whole set of schools has been divided in the same number of clusters, 3, and in the same numerical parts (11, 10, 20 elements), but the sets obtained are different for their composition and for the identified reference building, too.
- With K-medoids it was possible to obtain clusters with a higher Adjusted Index of Determination R^2_{adj}.
- With K-medoids, centro-types are identified in real objects included in the initial dataset, which is really useful to find a representative building.
- With K-medoids, standard deviation and the sum of square Euclidean distances result to be lower, confirming the more compactness and homogeneity of K-medoids models.
- Predictive models obtained with K-medoids generally result to fit better.

Acknowledgement

The authors would like to thank the Province of Treviso (Provincia di Treviso) for having made the schools database available for this research. The present study has been funded by the project "Klimahouse and Energy Production" in the framework of the programmatic-financial agreement with the Autonomous Province of Bozen-Bolzano of Research Capacity Building.

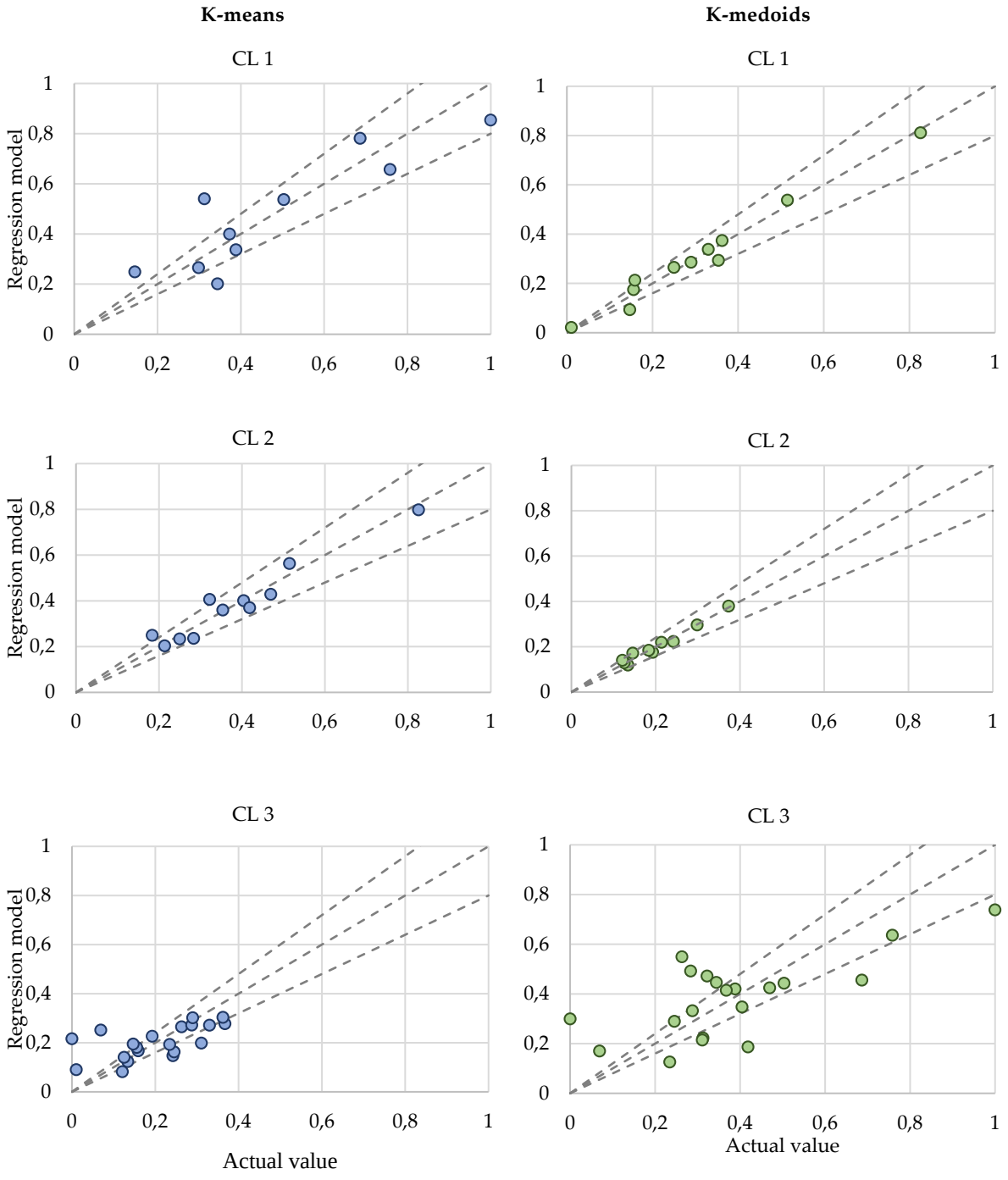

Fig. 4 - Estimated vs actual slopes of energy signatures of the schools in the three clusters. K-means clusters in blue colour, K-medoids ones in green colour. The dashed lines indicate a deviation of ±20 %

References

Arambula Lara, R., G. Pernigotto, F. Cappelletti, P. Romagnoni, A. Gasparella. 2015. "Energy audit of schools by means of cluster analysis". *Energy and Buildings* 95: 160-171. doi:10.1016/j.enbuild.2015.03.036.

Androulakis, N.D., K.G. Armen, D.A. Bozis, K.T. Papakostas. 2016. "Simulation of the thermal performance of a hybrid solar-assisted ground-source heat pump system in a school building". *International Journal of Sustainable Energy* 1-14. doi: 10.1080/14786451.2016.1261865.

Arora, P., S. Varshney. 2016. "Analysis of K-Means and K-Medoids Algorithm For Big Data". *Procedia Computer Science* 78: 507-512. doi: 10.1016/j.procs.2016.02.095.

CEN. 2008. *EN 15603 - Energy performance of buildings. Overall energy use and definition of energy ratings*. Brussels, Belgium: CEN.

European Parliament and Council of the European Union. 2012. "Directive 2012/27/EU, on energy efficiency, amending directives2009/125/EC and 2010/30/EU and repealing Directives2004/8/EC and 2006/32/EC". *Official Journal of the European Union* L315/1, 14 11.

Junjie, W. 2012. *Advances in K-means clustering. A data mining thinking*. Springer Thesis, Springer.

Kaufman, L., P.J. Rousseeuw. 2005. *Finding groups in data. An introduction to cluster analysis*. Wiley.

Niemelä, T., R. Kosonen, J. Jokisalo. 2016. "Cost-optimal energy performance renovation measures of educational buildings in cold climate". *Applied Energy* 183: 1005-1020. doi:10.1016/j.apenergy.2016.09.044.

Park, H., J. Lee, C. Jun. 2009. "A simple and fast algorithm for K-medoids clustering". *Expert Systems with Applications* 36(2): 3336-3341. doi10.1016/j.eswa.2008.01.039.

Pistore, L., G. Pernigotto, F. Cappelletti, P. Romagnoni, A. Gasparella. 2016. "From energy signature to cluster analysis: an integrated approach". In: *Proceedings of the 4th International High Performance Buildings Conference*. West Lafayette, U.S.A.: Purdue University.

Reynolds, A. P., G. Richards, V.J. Rayward-Smith. 2004. The application of K-medoids and PAM to the clusteing of rules. In *Intelligent Data Engineering and Automated Learning – IDEAL 2004. Lecture Notes in Computer Science* 3177.

Stavrakakis, G.M., A.V. Androutsopoulos, J. Vyörykkä. 2016. "Experimental and numerical assessment of cool-roof impact on thermal and energy performance of a school building in Greece". *Energy and Buildings* 130: 64-84. doi:10.1016/j.enbuild.2016.08.047