

## Using principal components analysis and clustering technics to define typical buildings

C. Lehmann

*Génie Civil et Mécanique, University of La Rochelle, France*

N. Gaitani

*Department of Environment & Natural Resources Management, University of Ioannina, Greece*

M. Santamouris

*Faculty of Physics, Department of Applied Physics, University of Athens, Greece*

### ABSTRACT

The present paper presents a method to characterize the typical building from a group by applying principal components analysis (PCA).

The method has been developed on a sample of secondary education school buildings in Greece. The purpose is to define the typical building in order to propose generalized improvements for energy efficiency of the building stock concerned. Therefore seven variables from questionnaires have been analyzed: heated surface, age of the building, insulation of the building, number of classrooms, number of students, school's operating hours and age of the heating system. Considering the frequency distribution of the examined variables, the typical building is defined as the closest to the sample median.

As a first step, a principal components analysis has been applied for transforming the original interrelated variables into the same number of new, uncorrelated variables called the principal components in order to consider the load of the variables and to reduce the dimension of the multivariate problem. Then, in the principal components coordinate system, the typical school has been identified as the closest to the sample median using Euclidean distance. Furthermore, k-means clustering technique may be applied to classify buildings significantly for a more extensive analysis of the stock.

The principal components coordinate system has eased the analysis of the multivariate sample and its classification and could as well provide a significant two-dimensional graphic of the sample.

Keywords: Energy classification, PCA, typical building, energy audit, energy efficiency in school buildings, clustering.

### 1. INTRODUCTION

To define the typical school in term of heating load out of an energy class based on oil consumption for heating, seven variables have been selected:

- Heated surface
- Age of the building
- Insulation of the building (dichotomous variable: 0 for non insulated, 1 for insulated)
- Number of classrooms
- Number of students
- School's operating hours per day (discrete variable)
- Age of the heating system

Considering the differences between the frequency distribution of each variable and the nature of the variables (continuous, discrete and dichotomique), the typical school is better defined with the median than with the mode or the mean. Thus, according to the data given, the typical school is defined as the closest for the seven variables to the median values; The closeness is measured with Euclidean distance.

### 2. USING PRINCIPAL COMPONENTS ANALYSIS

#### 2.1 The principal components analysis

Principal components analysis (PCA) is per-

formed in order to simplify the description of a set of interrelated variables by reducing the dimensionality of the multivariate problem. The technique (Afifi, 1996) can be summarized as a method for transforming the original variables into the same number of new, uncorrelated variables called principal components. Each principal component (PC) is a linear combination of the original variables and one measure of the amount of information conveyed by each PC is its variance. The principal components are arranged in order of decreasing variance, thus the most informative PC is the first and the least informative is the last.

Practically, an investigator may wish to reduce the dimensionality of a problem (i.e. reduce the number of variables without losing much of information), this can be achieved by choosing to analyse only the first PCs; the number of components selected may be determined by examining the proportion of total variance explained by each component; the PCs not analysed convey only a small amount of information since their variances are small. Thus instead of analysing a large number of original intercorrelated variables, the investigator can analyse a small number of uncorrelated PCs. Moreover, by examining the coefficients of the linear combination between the PCs and the original variables, a physical interpretation may be given to the first principal components as “supervariables”.

### *2.2 Define the typical school in the principal components coordinate system*

Among the seven variables, some should be interrelated with different correlation factors. Then, using the principal components as new variables and describing the data in this new coordinate system, should be more meaningful to point out the typical school by the fact that the correlation and the load of the variables are considered. Thus, the definition of the typical school becomes the closest school to the median point in this new coordinate system.

Therefore, the Euclidian distance performed in the seven dimensional PCs coordinate system is not invariant to changes in variance among the PCs, the accuracy of the method is achieved substantially from the fact that the variance decreases with the amount of information conveyed.

Furthermore, by plotting the data in the bidimensional first two PCs coordinate system, subgroups of schools may appear graphically as different regions; thus the distinction between these subgroups, expressed by the two first principal components and their physical interpretations, displays different types of schools.

## 3. PREPARATION OF THE DATA

Principal components analysis is sensitive to outliers (Barbara, 1996); As a prior preparation of the data, outliers have been identified and removed (Tuckey rule for continuous variables and a multiple of standard deviation from the median for the discrete variable). Then, each variable has been standardized by subtracting the mean and dividing by the standard deviation to perform principal components analysis on dimensionless and identical variance variables. Finally, the schools with missing data have been screened: The identification of the typical school (procedure explained subsequently) has been performed in different cases and by performing this test, it has been decided to remove from the sample the schools with at least one missing data.

## 4. EXAMPLES OF IDENTIFICATION OF TYPICAL SCHOOL

The typical school selection is performed on two different samples in order to illustrate the kind of results obtained. The PCA has been performed with the Matlab function PRINCOMP.

### *4.1 Sample A*

Sample A contains 338 schools (429 schools observed minus 91 schools with outliers or missing data).

#### *4.1.1 Principal components analysis*

The following correlation matrix shows that PC1 is mostly a linear combination of the variables age of the building, insulation of the building and age of the heating system. PC1 increases with the age of building and the age of the heating system whereas it decreases with the variable insulation of the building, indeed old buildings are not insulated. The new variable PC1 resumes the informations regarding the age. PC2 is mostly a combination of the vari-

PC1	-0.02	0.61	-0.55	-0.05	-0.04	-0.08	0.56	heated surface
PC2	-0.50	-0.06	0.09	-0.61	-0.57	-0.21	0.01	age of the building
PC3	0.35	-0.03	0.11	0.05	0.00	-0.93	0.01	insulation of building
PC4	-0.76	0.00	0.05	0.20	0.56	-0.27	0.04	number of classrooms
PC5	-0.12	0.04	-0.64	0.19	-0.17	-0.12	-0.70	number of students
PC6	0.20	-0.11	-0.25	-0.73	0.57	0.00	-0.15	school operating hours
PC7	0.02	0.78	0.44	-0.12	0.08	0.02	-0.42	age of the heating system

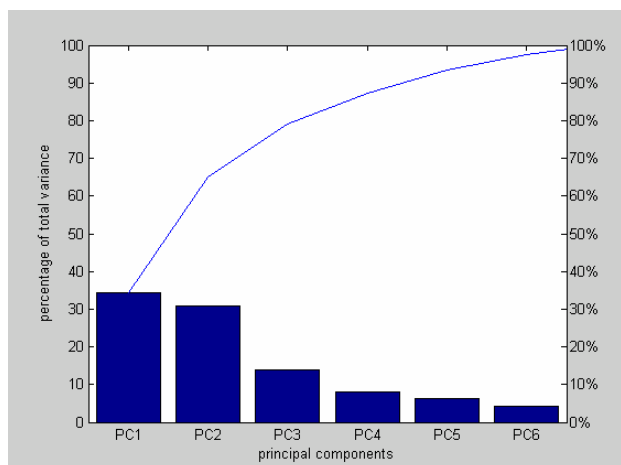


Figure 1: Percentage of total variance of each PC.

ables heated surface, number of classrooms and number of students (all the loading coefficients have the same sign), PC2 includes the information regarding the size. PC3 mostly represents the school operating hours.

The first two PCs convey a large amount of information quantified by 65% of cumulative percentage of total variance (Fig. 1).

#### 4.1.2 Selection of the typical school

The typical school is the closest school to the median point whose coordinates are the medians of each PC. The measure of closeness is the Euclidean distance in the 7 PCs coordinate system.

Table 1 displays the characteristics of the typical school selected compared to the median and the standart deviation of the sample for each category, whereas Table 2 displays the selection of the ten closest schools to the median point.

As shown on the Figure 2, the first two PCs do not convey complete information.

#### 4.1.3 Validation of the method using PCA to select typical building

In order to validate the selection of the ten most typical schools among the sample, another technique has been employed based on a “funnel” type selection: each “funnel” is defined as a cer-

Table 1: Characteristics of the typical school.

	School N°1174	Median	Standart deviation
Heated surface (m <sup>2</sup> )	2150	1835	1652
Age of building (years)	14	18	15
Insulation of building	1	1	
Number of classrooms	17	17	11
Number of students	206	230	208
School operating hours	6	6	1
Age of heating system (years)	14	15	11

Table 2: Selection of the ten closest schools.

School selection	N° 1	N° 2	N° 3	N° 4	N° 5
School's ID number	1174	1101	448	183	885
School selection	N° 6	N° 7	N° 8	N° 9	N° 10
School's ID number	656	590	212	287	647

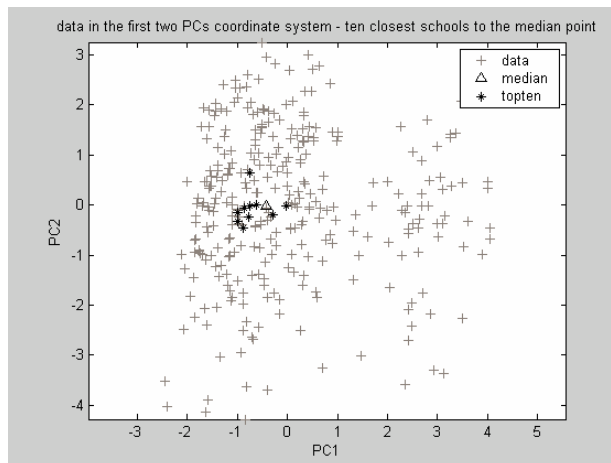


Figure 2: Data in the first two PCs coordinate system.

tain aperture around the median value of the corresponding original variable, the schools which pass through all seven funnels are selected as typical. Thus, for each category, the value of a school selected is “close” to the median.

Funnel description: For each variable, the funnel is designed to select the schools with a value between an inferior limit  $Q_{50-x}$ , the (50-x)th quintile, and a superior limit  $Q_{50+x}$ , the (50+x)th quintile, so each funnel is designed according to the median value and the distribution

Table 3: Selection of the schools with funnels.

School's ID number	Between	Between	Between
	Q <sub>40</sub> & Q <sub>60</sub>	Q <sub>35</sub> & Q <sub>65</sub>	Q <sub>30</sub> & Q <sub>70</sub>
			212
		269	269
			287
			319
			327
			474
			590
			647
		656	656
			681
		885	885
		1101	1101
			1122
	1174	1174	1174

of the category. At the end of the selection, typical schools have passed through all the funnels and their number depends of the value  $x$  which defines the operture. For the dichotomous variable insulation of the building, the funnel has been replaced by a selection of only the insulated building (most frequent).

Most of the schools selected (Table 3) using this method appear also in the selection using distance to median point in the PCs coordinate system. The “funnels” method, robust by its simplicity, does not give a sorted selection of the schools, but the comparison between the results could validate the method of using PCA to select a typical building.

#### 4.2 Sample B

Sample B contains 247 schools (319 schools observed minus 72 schools with outliers or missing data).

##### 4.2.1 Principal components analysis

With this sample, the loading coefficients of the first PCs in the following correlation matrix are not well separated; The physical interpretation of the first PCs will be discussed subsequently.

$$\begin{pmatrix} PC1 \\ PC2 \\ PC3 \\ PC4 \\ PC5 \\ PC6 \\ PC7 \end{pmatrix} \begin{pmatrix} 0.41-0.47 & 0.41 & 0.43 & 0.39 & 0.04-0.33 \\ -0.36-0.40 & 0.42-0.42 & -0.43 & -0.03 & -0.41 \\ 0.02-0.01 & 0.03 & 0.04 & 0.01 & -1.00 & 0.03 \\ -0.33 & 0.14 & -0.49 & 0.21 & 0.17 & -0.04 & -0.74 \\ 0.70 & 0.14 & -0.19 & -0.02 & -0.61 & -0.01 & -0.30 \\ 0.32 & 0.05 & -0.09 & -0.77 & 0.51 & -0.03 & -0.18 \\ -0.03 & 0.76 & 0.61 & 0.06 & 0.06 & 0.01 & -0.22 \end{pmatrix} \times \begin{pmatrix} heated\ surface \\ age\ of\ the\ building \\ insulation\ of\ building \\ number\ of\ classrooms \\ number\ of\ students \\ school\ operating\ hours \\ age\ of\ the\ heating\ system \end{pmatrix}$$

As the sample A, the first two PCs convey a

Table 4: Characteristics of the typical school.

	School N°302	Median	Standart deviation
Heated surface (m <sup>2</sup> )	1400	1260	1489
Age of building (years)	21	21	18
Insulation of building	1	1	
Number of classrooms	14	16	11
Number of students	176	191	259
School operating hours	6	6	1
Age of heating system (years)	21	18	10

Table 5: Selection of the ten closest schools.

School selection	N° 1	N° 2	N° 3	N° 4	N° 5
School's ID number	302	661	1131	828	983
School selection	N° 6	N° 7	N° 8	N° 9	N° 10
School's ID number	892	655	989	397	72

large amount of information quantified by a cumulative percentage of total variance equal to 65,5%.

##### 4.2.2 Selection of the typical school

Table 4 displays the characteristics of the typical school selected whereas the selection of the ten closest schools to the median point is presented Table 5.

##### 4.2.3 Comparison with the “funnel” selection

Most of the schools selected (Table 6) using this method appear also in the selection using distance to median point in the principal components coordinate system. The comparison between the results validates once more the method of using PC analysis to select typical building.

## 5. FURTHERMORE SUB-GROUPING

### 5.1 Physical interpretation of the principal components

Depending on the original variables and the sample, physical interpretation may be given to the first components as “supervariables”, inter-

Table 6: Selection of the schools with funnels.

	Between Q <sub>40</sub> & Q <sub>60</sub>	Between Q <sub>35</sub> & Q <sub>65</sub>	Between Q <sub>30</sub> & Q <sub>70</sub>
School's ID number	302	302	114
		397	302
		661	397
			661
			828
		892	892
		983	892
			983
		1131	1131

pretation deducted from the loading coefficients given in the correlation array. If the coefficients of one principal component are very distinct between large loadings and small loadings, then the principal component is a linear combination of only few original variables with maybe an underlying physical interpretation. If an interpretation can be given it would concern only the first PCs which convey the larger amount of information. Thus, for the particular sample A, the first PC resumes the information regarding the age whereas the second PC informs about the size of the school. Generally, the loading coefficients for each of the first PCs are not well separated as for the sample B but still an underlying physical interpretation may be given. Therefore an orthogonal rotation of the PCs can be performed with a parameterization in order to increase the high loadings and to decrease the small loadings coefficients of the first PCs.

To avoid an orthogonal rotation analysis of the PCs, a biplot popularised by Gabriel (Jolliffe, 1993), gives a display of the loadings for the first two PCs which may help its physi-

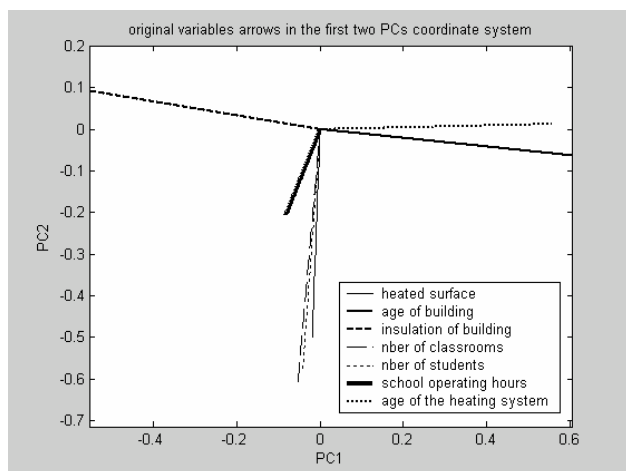


Figure 3: Original variables arrows in the first two PCs coordinate system.

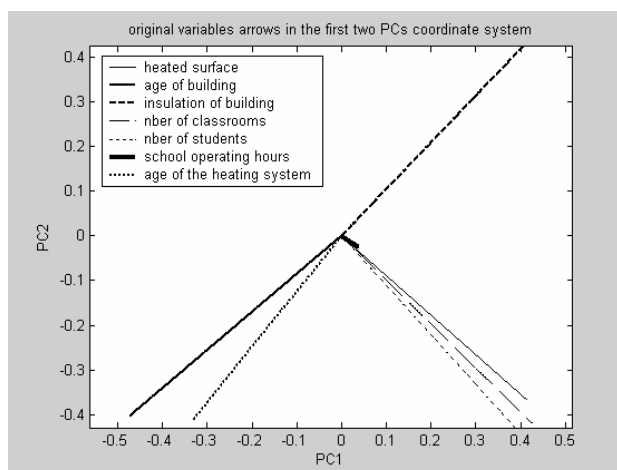


Figure 4: Original variables arrows in the first two PCs coordinate system.

cal interpretation: each original variable is denoted by its unit vector in the first two PCs coordinate system. Then, the angles between the arrows give an indication of the correlation between the original variables; Variables with arrows in similar directions tend to be positively correlated, whereas negative correlation leads to arrows in opposite directions; Uncorrelated variables tend to have arrows at right angles to each other; The degree of approximations and tendencies of this interpretation decreases when the cumulative percentage of total variance of the first two PCs increases.

### 5.1.1 Sample A

As noticed in the §4.1.1, a rotation of the PCs is not needed to give a physical interpretation of the first two principal components.

### 5.1.2 Sample B

The Figure 4 illustrates that the physical interpretation given to the first two PCs by rotating them could be the same as the sample A.

As illustrated, the biplot of the original variables in the first two PCs coordinate system may guide the investigator to establish the correlation among the original variables and to propose a physical interpretation of the first two PCs without rotating them.

### 5.2 Sub-groups of a sample

Sub-groups of schools may appear from differences among the categories. We could split the group of schools regarding one original variable, but defining sub-groups of schools by analysing the data expressed in the PCs coordinate

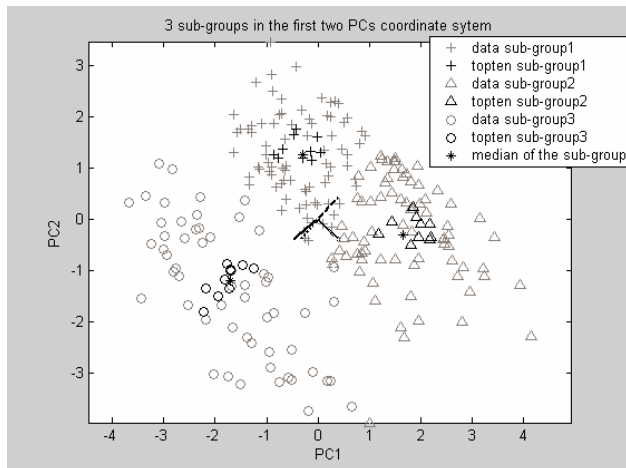


Figure 5: Original variables arrows in the first two PCs coordinate system.

system is a way to include several variables in the splitting process. The physical interpretation of the first PCs gives then the meaning which drives the differentiation among the sub-groups. If meaningful sub-groups appear, then a typical school for each subgroup can be identified as well.

#### *Sample B – Example of 3 sub-groups*

The K-means clustering technique has been applied on the data defined in the PCs coordinate system to highlight three sub-groups (Fig. 5).

The biplot of both observations and variables conveys useful information: Observations on the same side as the arrow will tend to take larger value than the average on that variable, whereas smaller value than the average may concern observations on the opposite side of the arrow; The closer an observation is to the direction of the arrow, and the further it is from the origin, the larger is likely to be the value of the observation on that variable.

The schools in the sub-group 3 are older than for the sub-groups 1 and 2: The age of the building and the age of the heating system are higher; All the schools in the sub-group 3 are not insulated whereas they are all insulated in the sub-groups 1 and 2. The schools in the sub-group 2 have a bigger size than for the sub-group 1: The heated surface and the numbers of classrooms and students are higher.

## 6.CONCLUSION

The selection of the typical school among a sample as the closest to the medians in the prin-

cipal components' coordinate system is an accurate multivariate statistical method; The principal components convey the load of each variable and the Euclidean distance which measures the closeness gains to be performed in the PCs coordinate system because of the decreasing of variance with the amount of information conveyed. Furthermore, by reducing the dimensionality of the problem, a bidimensional graphic in the first two PCs coordinate system may help the investigator to understand the correlation between variables and defining sub-groups as well.

## REFERENCES

- Afifi, A.A., 1996. Computer-aided multivariate analysis. London: Chapman & Hall.
- Barbara, G., 1996. Using multivariate statistics. New York: HarperCollins College Publishers.
- Jolliffe, I.T., 1993. Principal component analysis: A beginner's guide – II. Pitfalls, myths and extensions. *Weather* 48: pp 246-253.