Multivariate Methods for Evaluating Building Energy Efficiency

Thomas Olofsson, Staffan Andersson and Jan-Ulric Sjögren, Umeå University

ABSTRACT

For an owner or operator of a building, benchmarking can be a useful guide for finding out how energy efficient the building is and identifying what to improve. For successful evaluation of the building energy efficiency, the categorization as well as the parameter identification has decisive importance. That selection can be based on mathematical modeling such as linear regression accompanied with more or less user expert knowledge. The selection, however, is not a simple task since analyses based on statistical data are sensitive to correlations between different measured parameters. For improving that analysis multivariate methods such as Principal Component Analysis (PCA) can be a valuable support.

We demonstrate here how PCA can be a useful tool for investigating aggregated statistical datasets. The investigation illustrates how a set of building performance parameters exhibits different relations depending on how the categorization is made, which is relevant to consider when benchmarking. The study is based on a national Swedish database of aggregated energy use and building performance statistics.

Introduction

Since the days of the energy crisis, 30 years ago, major efforts have been focused on reducing energy dependency and promoting sustainable alternatives. Successful innovations accompanied by long-term environmental legislation and strategically economical subsidies for sustainable buildings have in many ways promoted good conditions to reduce energy use. Thus, the use of energy in the building sector, for example in Sweden, still has the same magnitude as 30 years ago, in spite of the fact that the total building area has increased (Swedish National Energy Administration 2004).

Although successful work has been done more recently, the energy use in buildings has to be considerably reduced in order to meet demands of a future sustainable society. In this perspective relevant questions are, for example: How large is the technical energy efficiency potential for individual buildings, and for the total building stock? How could savings be identified? Benchmarking can be a guide to provide some answers.

Benchmarking building energy efficiency requires large datasets of statistical representative information. Such datasets can be based on simulations or aggregated statistics. Simulations reveal the ideal behavior of a building or its behavior under standardized weather and operating conditions, however they ignore aspects such as operation and maintenance. For valid benchmarking, measured actual performance data are needed; simulated data have been found to be misleading in many cases, see e.g. (Nilsson 2003). In the literature several examples of benchmarking can be found, see for example (Hicks and Clough 1998: Zmeureanu et al. 1999; Laustsen 2001; Richalet et al. 2001; Federspiel et al 2002; http://poet.lbl.gov/arch/).

Each benchmarking procedure is developed for answering certain questions using a defined set of data. There are different procedures to identify conditions on comparable data sets. That identification can be based on expert knowledge and/or mathematical or statistical

investigations. An example is stepwise linear regression modeling, which was found usable to identify the strongest causes to the energy use in office buildings (Sharp 1996). This is however not a simple task since analyses based on statistical data are sensitive to correlations between different measured parameters. Principal component analysis (PCA), used in this work, is a statistical method that can be of help for selecting datasets for benchmarking building energy efficiency.

Data

In an earlier project entitled 'e-nyckeln' (Vitec Fastighetssystem AB 2003) data were collected for commercial and residential buildings in Sweden. The main aims of that project were to develop a method for collecting data for classification and energy consumption, and a tool for benchmarking buildings over the internet (http://www.vitec.se/enyckeln/index.htm).

In the methodology developed, the property holder provided the building classification information and allowed the project team to retrieve monthly energy data from the building energy commissioning software. All the information was entered into a database having an SQL format. The collected classification data include about 80 parameters, such as category of tenants, geometry of the building, HVAC-system, cooling system, control strategy, climate-zone, etc. For each registered building, monthly data included cold and hot water consumption, energy for cooling and heating, and actual climate data. Energy data typically covered one to three years. The process to collect data started in 2002 and is still running. For the moment there are about 500 documented buildings in the database.

Generally, the collected energy-use data do not include the total supplied energy, but only the energy paid by the real estate owner or operator, i.e. energy for heating and cooling and hot water preparation. In Sweden the energy supplier usually charges tenants for energy of appliances, etc. Hence, for this particular project, the tenants' bills were not available. For a thorough analysis, the total energy use is essential. Because of this data limitation, the results are discussed in the context of introducing the methodology based on PCA as opposed to establishing the appropriateness of the dataset for benchmarking specific buildings.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) can be described as a multivariate procedure, which rotates the axis to maximize the variance or description. Essentially, a set of correlated variables is transformed into a new set of uncorrelated variables that are ordered by reducing variance. The new uncorrelated variables, or Principal Components (PC), are linear combinations of the original variables, and the last of these variables can be removed with minimum loss of real data. PCA can thus be used to reduce the dimensionality of a data set while retaining as much information as is possible.

Suppose a p-dimensional variable in the format $\boldsymbol{X}^T = (X_1, X_2, \ldots, X_p)$ with mean μ and covariance matrix Σ. The transformation of $\boldsymbol{X}^T$ can be described as an orthogonal rotation (Chatfield & Collins 92). The new set of variables $\boldsymbol{Y}^T = (Y_1, Y_2, \ldots, Y_p)$ is linear combinations of the Xs according to

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p = \mathbf{a}_j^T\mathbf{X} \qquad (1)$$

where the vector of constants $\mathbf{a}_j^T = (a_{1j},a_{2j},...,a_{pj})$ is orthonormal, i.e. $\mathbf{a}_j^T\mathbf{a}_j = 1$. The principal components, Y_j, are uncorrelated with decreasing variance. Thus, the first principal component Y_1, or PC1, can be found by choosing an $\mathbf{a}_1$ that maximizes the variance of Y_j. Using equation (1) and by introducing Σ the variance can be given by

$$Var\ (Y_1) = Var\ (\mathbf{a}_1^T X) = \mathbf{a}_1^T\ \Sigma\ \mathbf{a}_1 \qquad (2)$$

Further, it can be assumed that Σ has p eigenvalues, or latent roots, which are defined according to $\lambda_1 > \lambda_2 > ... > \lambda_p \geq 0$. By applying the method of Lagrange multipliers it can be found that $(\Sigma - \lambda_j)\mathbf{a}_j = 0$. Thus, equation (2) can be written as

$$Var(Y_1) = \mathbf{a}_1^T(\lambda_1\mathbf{a}_1) = \lambda_1 \qquad (2)$$

The second principal component, PC2, with an eigenvalue equal to λ_2, is an orthogonal vector to PC1. PC2 is found with the same method and the procedure can be repeated until the last orthogonal PC is defined. A scaled vector representing the properties of the investigated parameter with respect to the principal components is given by the component loading defined as $\mathbf{a}_j$*. It is possible to obtain $\mathbf{a}_j$* by scaling $\mathbf{a}_j$ according to

$$\mathbf{a}_j^* = \sqrt{\lambda_j}\,\mathbf{a}_j \qquad (4)$$

If the component loadings are added in a matrix $\mathbf{C} = (\mathbf{a}_1^*,\mathbf{a}_2^*,..,\mathbf{a}_p^*)$ the definition of Σ is given by $\Sigma = \mathbf{C}\mathbf{C}^T$.

Multivariate methods such as PCA have not been used to any large extent in evaluations of building energy efficiency based on aggregated data. The method has been applied in related fields for guidance on how to reduce the number of modeled physical variables and also to describe the most significant explanation in a fewer number of transformed PCs, see for example (Ruch 1993; Reddy 1995; Olofsson 1998; del Barriao 2004).

Model

In the database, parameters of documented energy consumption and factors related to the technical performance and the operation of the facilities were available for the study. From a preliminary analysis, the following seven parameters were selected for further investigation:

In order to investigate the dataset a PCA was conducted to minimize the dimensionality. Selecting the most significant PCs, i.e. those with larger eigenvalues than the noise limit, set at 12.5%, the reduction in dimensionality was made. The component loadings from the significant PCs were used to indicate if certain parameters could be clustered and to determine if the clusters were related to some specific feature. The analysis of clusters was also supported by an investigation of correlation in the original parameters.

Table 1. Selected Parameters for the Investigation

Parameter	Data	Analyzed as
x1	The year of construction	The year A.D
x2	Heated building area	m^2
x3	HVAC-system	1- Natural ventilation 2 -Exhaust fan 3 - Supply and exhaust fan 4 – Supply and exhaust fan and heat recovery
x4	Strategy for running the system	1 - Continuous 2 - Part time 3 - Based on load
x5	The age of the control system	1 - Less than 5 years, 2 - 5-10 Years, 3 - Older than 10 years
x6	Occupancy sensors for controlling lighting	0 - No 1 - Yes
X7	Supplied district heating,	Energy Unit Index (EUI) kWh/m^2, year, degree-day corrected

Results and Discussion

Investigation on all Buildings with District Heating

The investigation was first conducted on all available buildings that were heated with district heating, representing a greater part of the buildings in the database. The data included about 280 buildings with district heating. The majority of those buildings were commercial buildings, e.g., offices, stores and schools but some of those were partially used for housing. The year the buildings were constructed ranged from 1899 to 2000 with a median at 1971. The heated floor area ranged from 168 to 37,200 m^2 (1,808 to 400,431 ft^2), with a median of 6,980 m^2 (75,135 ft^2). The EUI of supplied district heating ranged from 19 to 561 kWh/m^2 a year (1.77 to 52.12 kWh/ft^2 a year) with a median at 143 kWh/m^2 a year (13.28 kWh/ft^2 a year).

A PCA was made on the selected data set of seven parameters and the results are presented in Figure 1. The eigenvalues describe how large a part of the total variance is explained by each PC. In this work we have not considered PCs with eigenvalues smaller than 12.5%, i.e. the noise limit. The PCs are linear combinations of the original variables. This relation is explained in terms of the component loading. In the conducted analysis it is assumed that loading larger than 0.45 can be assumed as significant. The plots in Figures 1a and 1b illustrate the obtained eigenvalues and component loading.

The upper plot in Figure 1a illustrates the eigenvalue of each PC. The PCs with eigenvalues (L) over the limit of noise (Ln = 12.5%) can be assumed as significant. In the lower plot in Figure 1a, the first two components are plotted against the original variables x1 to x7. The component with the best explanation, PC1, is plotted as a solid line and the second component, PC2, as a dashed line. The component loadings of the first three principal components are illustrated in the histogram in Figure 1b. The eigenvalues of the third principal component in Figure 2 are lower than the noise limit 12.5% and can be neglected.

Figure 1a. The Values of the Eigenvalues, or Latent Roots, Minus a Defined Limit of Noise (12.5%) Are Presented for each PC in the Upper Plot, Where PC1 Is the First Root to the Left and PC7 Is the Last Root to the Right. In the Lower Plot the Component Loading Is Shown for Variable X1 to X7, from the Left to the Right on the X-Axis, Based on Data of Accessible Buildings with District Heating.

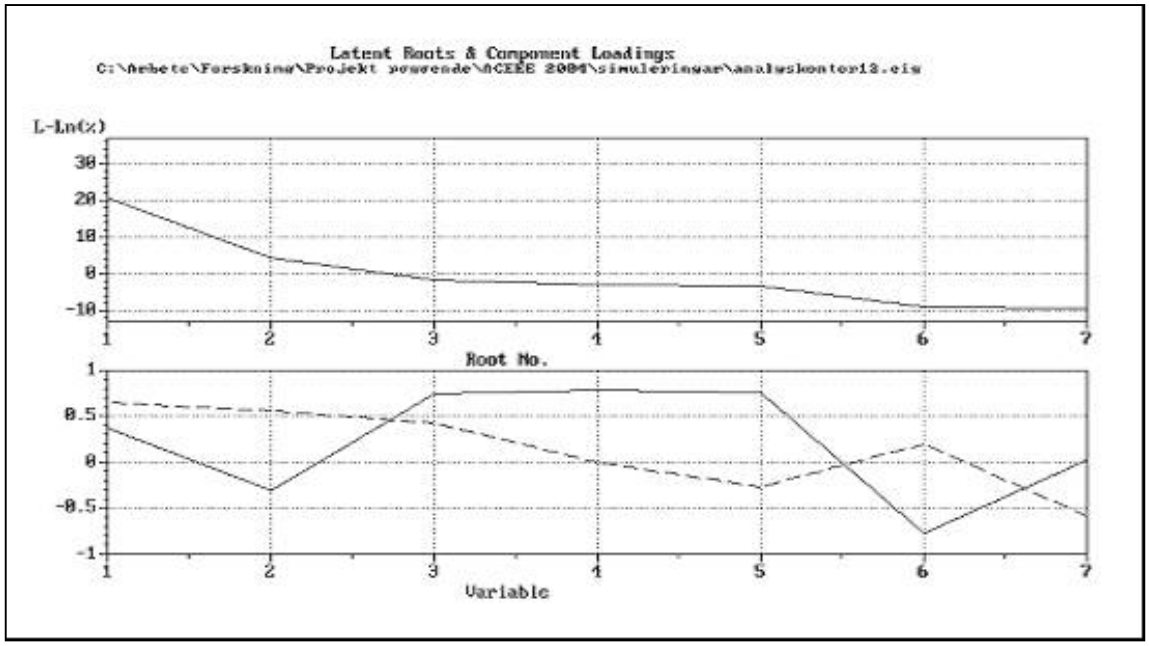
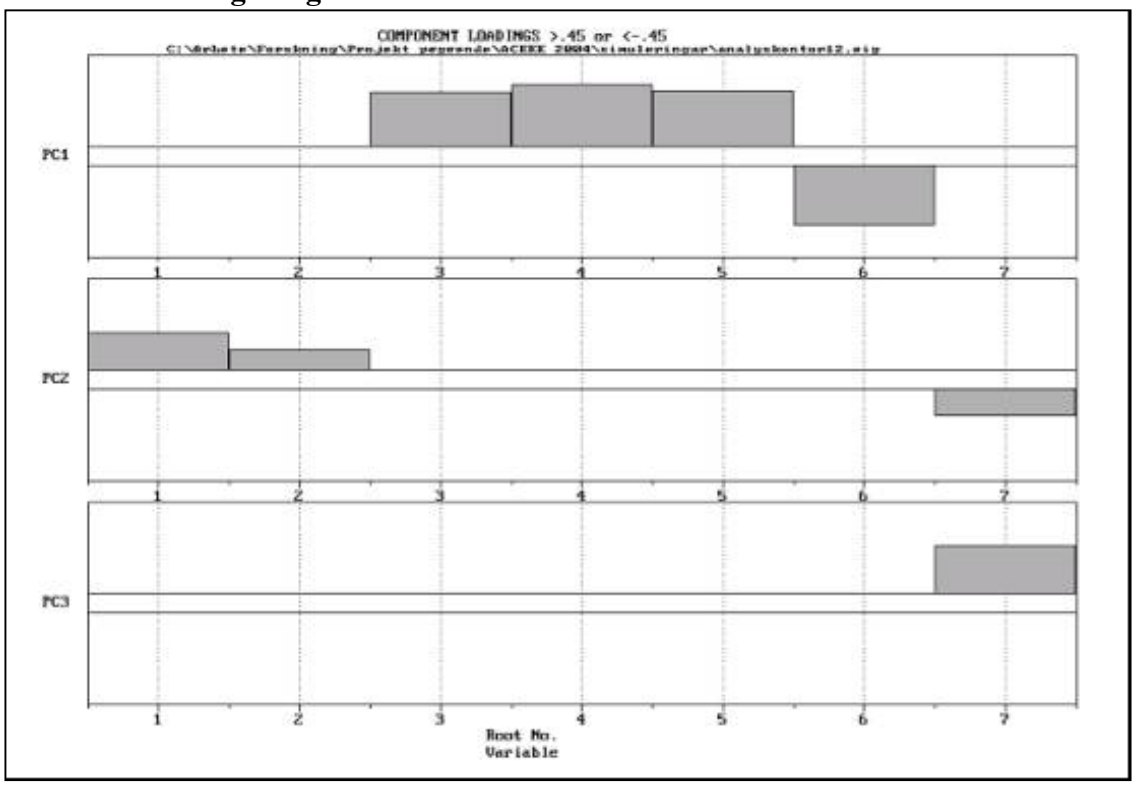


Figure 1b. Histograms Illustrating the Component Loadings for the Variable X1 to X7, from the Left to the Right on the X-Axis, of Accessible Buildings with District Heating. Loading Larger than 0.45 Is Illustrated for the Three First PCs.



The PCA on the 280 buildings gives two significant PCs. The first PC, PC1, was rather dependent on x3, x4, x5 and x6, i.e. the type of HVAC-system, control system, control strategy, and existence of occupancy sensors for lighting. The second PC, PC2, was dependent on the parameters x1, x2 and x7, i.e. age of the building, floor area and EUI.

A correlation analysis was also conducted on the dataset. In Table 2 the correlation matrix is presented. The correlations to EUI are not very significant for any variable. However, the correlations are relatively larger for the variables with stronger component loadings in PC1.

Table 2. The Correlation Matrix Calculated for the Investigated Seven Parameters, Of the Accessible Buildings with District Heating

	x1	x2	x3	x4	x5	x6	x7
X1	1						
X2	0.094	1					
X3	0.405	-0.053	1				
X4	0.113	-0.177	0.653	1			
X5	0.140	-0.259	0.289	0.386	1		
X6	-0.133	0.201	-0.34	-0.440	-0.664	1	
X7	-0.144	-0.102	-0.122	0.105	0.063	-0.015	1

The strongest obtained PC, PC1, which explains the largest part of the variance of the dataset, reveals dependency of HVAC-system, strategy for running the system, the age of the control system and occupancy sensors for controlling lighting. In the less significant PC, PC2, we detected dependency of EUI and the year of construction and heated building area.

For benchmarking of EUI it is desirable that EUI is included in the most significant PC. The weak dependency of EUI from this study can indicate that the dataset has to be selected in another way.

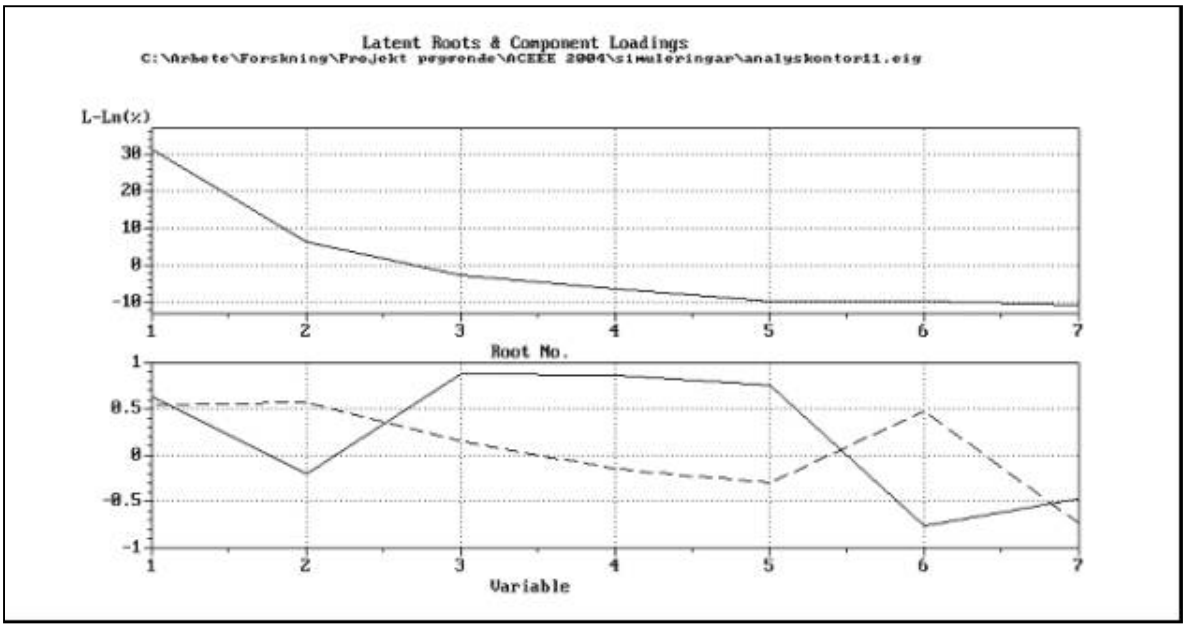
For the chosen dataset, buildings heated by district heating, the data included different categories of tenants, which might explain the results. Our next step was, therefore, to narrow the data set to office buildings with district heating and to conduct a PCA to investigate how that might change the results.

Investigation of Office Buildings with District Heating

A little less than 50 of the original 280 buildings had offices. The year of construction for these buildings ranged from 1903 to 1993 with a median at 1967. The heated floor area ranged from 250 to 34,200 m^2 (2,691 to 368,138 ft^2), with the median 4280 m^2 (46,071 ft^2). The EUI of supplied district heating ranged from 22 to 280 kWh/m^2 a year (2.04 to 26.01 kWh/ft^2 a year) with a median at 125 kWh/m^2 a year (11.61 kWh/ft^2 a year). It should be noted that since the set of office buildings with district heating was relatively small, all buildings of that category had to be included in the analysis. Thus, buildings with a minor part of offices are also included.

A PCA was then performed on the seven variables for the 50 office buildings. The results are shown in Figure 2.

Figure 2a. The Values of the Eigenvalues, or Latent Roots, Minus a Defined Limit of Noise (12.5%) Are Presented for each PC in the Upper Plot, where PC1 Is the First Root to the Left and PC7 Is the Last Root to the Right. In the Lower Plot the Component Loading Is Shown for Variable X1 to X7, from the Left to the Right on the X-Axis, Based on Data of Office Buildings with District Heating.



According to the upper plot in figure 2a the first two PCs, PC1 and PC2, were significant. The calculated component loadings in Figure 2b show that the first PC, PC1, is dependent on mainly x1, x3, x4, x5, x6 and x7, i.e. variables illustrating the age of the building, type of HVAC-system, control system and strategy, occupancy sensors for lighting and EUI. The second PC, PC2, is mainly dependent on the parameters x1, x2 and x7, i.e. age of the building, floor area, occupancy sensors and EUI.

A correlation matrix, from a multivariate analysis on the dataset, was also calculated, see Table 3. The correlation was rather significant for the variables x3, x4, x5 and x6, as in Table 2, but in this matrix also between x7 and x1. That is in fair agreement with the results of the PCA.

The results in Figure 2 show the same strong dependency of PC1 to HVAC-system, control system and control strategy and occupancy sensors for lighting, but also to EUI. It can also be noticed that the eigenvalue of PC1 is larger and thus more significant and that the component loadings have become larger. For PC2 similar results were obtained as for the previous data set. The categorization of the first data set to office buildings was thus found to be desirable for benchmarking EUI, since EUI was better described in the variance of that data set.

This simple exercise illustrates the importance of selecting data since the selection evidently influences the appropriateness of using the data set for modeling/benchmarking a particular variable (in this case EUI).

Figure 2b. Histograms Illustrating the Component Loadings for the Variable X1 to X7, from the Left to the Right on the X-Axis, of Office Buildings with District Heating. Loading Larger than 0.45 Is Illustrated for the Three First PCs.

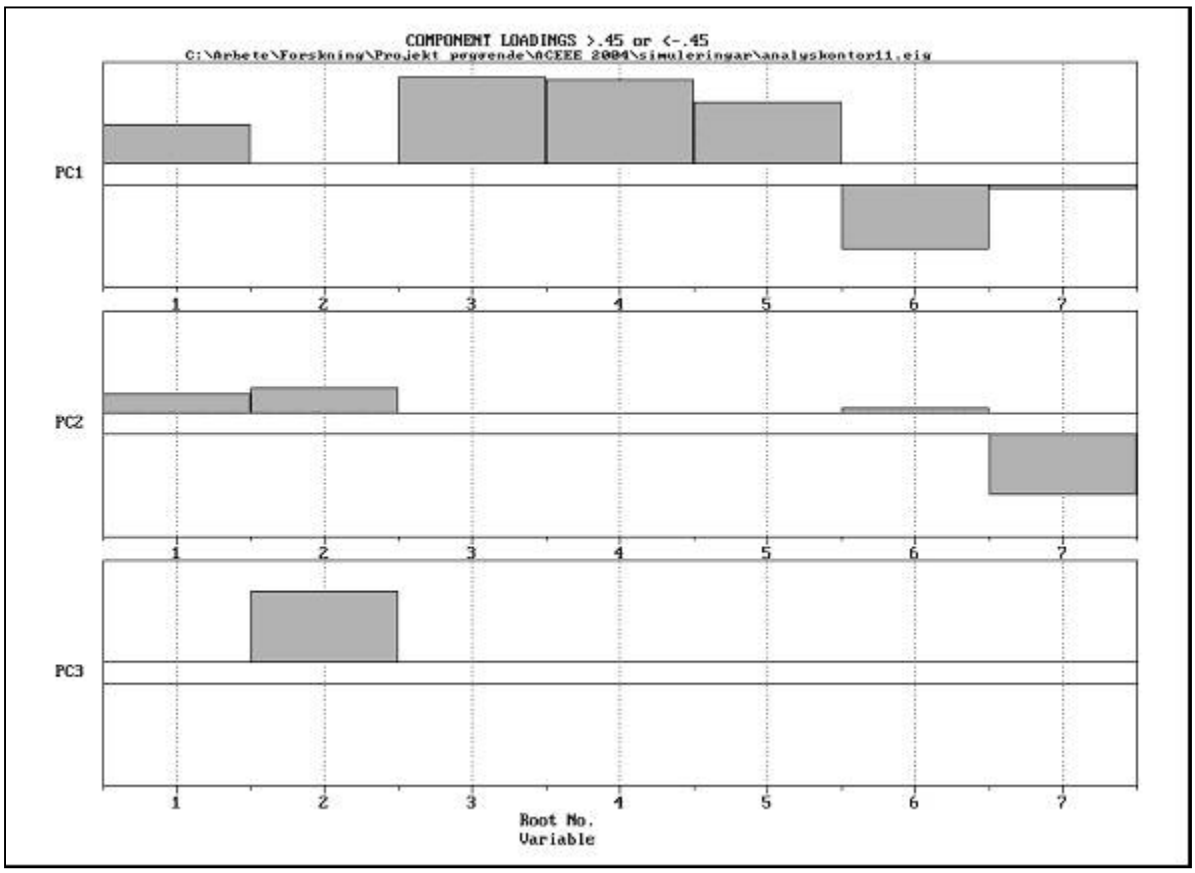


Table 3. The Correlation Matrix Calculated for the Investigated Seven Parameters, Of Office Buildings with District Heating

	x1	x2	x3	x4	x5	x6	x7
X1	1						
X2	0.020	1					
X3	0.513	-0.072	1				
X4	0.331	-0.120	0.766	1			
X5	0.371	-0.137	0.480	0.602	1		
X6	-0.224	0.331	-0.541	-0.656	-0.649	1	
X7	-0.615	-0.110	-0.490	-0.238	-0.064	0.050	1

General Comments on the Results

For a successful analysis of this kind the need for descriptive parameters and lack of parameters can always be discussed. For example it would have been an advantage with a parameter indicating the intensity of the activities in the buildings, indoor temperature and outdoor airflow, as well as the electricity use of e.g., computers, office equipment, lighting, etc. Such parameters were not available for this study.

All office buildings in this investigation were found in a distance of 250 km (155 miles) from Stockholm, thus the differences in climate were found to have a minor influence. For a larger study categorization into climate zones may have been necessary.

Summary

Benchmarking building energy efficiency has gained a growing interest, since for example in the European Union, classification probably will be compulsory (Council of the EU 2001). Thus, applicable methodologies to benchmark will be needed. That will call for reliable procedures to find indications of how to categorize data for benchmarking purposes.

A result of this study, which can have importance for categorization in particular, and benchmarking in general, is that the choice of reference dataset will have a decisive importance for the results. Thus, for obtaining reliable validations we advise carefulness when choosing data.

In this study PCA was used as a support tool for the identification of a suitable dataset. There are also other multivariate methods, such as partial least squares (PLS) that can be useful for the next step that involves modeling and prediction.

Acknowledgement

Erik Fällman, Stig Byström and Stefan Berglund at the department of Applied Physics and Electronics, Umeå University, are gratefully acknowledged for preparing the investigated database.

The authors would like to thank The Swedish Construction Sector Innovation Center (BIC), The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) and NCC AB for their financial support of this work.

References

del Barriao E.P. and G. Guyon 2004, Application of parameters space analysis tools for empirical model validation, *Energy and Buildings*, Vol. 36, pp. 23-33.

Chatfield C. and A. J. Collins 1992, *Introduction to multivariate analysis*, Chapman and Hall, London.

Council of the EU 2001, Interinstitutional file: 2001/0090 (COD).

Federspiel C., Zhang Q. and Arens E. 2002, Model-based Benchmarking with Application to Laboratory Building, *Energy and Buildings*, Vol. 34, pp. 203-214.

Hicks T. and Clough D. 1998, Energy Star ® Building label: Building Performance through Benchmarking and Recognition, *Proceedings of the 1998 ACEEE Summer Study of Energy Efficiency in Buildings. American Council for an Energy-Efficient Economy*, Washington DC, Vol. 4, pp. 4.205-4.210.

Laustsen J. 2001, Mandatory labeling of buildings: The Danish experience, *Sustainable Buildings*, vol. 4, pp. 12-14.

Lawrence Berkeley National Laboratories 2004, http://www.lbl.gov/arch

Nilsson A. 2003, *Energianvändning i nybyggda flerbostadshus på Bo01-området i Malmö*, TVBH-3045 (in Swedish).

Olofsson T., Andersson S. and Östin R. 1998, Using CO2 Concentrations to Predict Energy Consumption in Homes, *Proceedings of the 1998 ACEEE Summer Study of Energy Efficiency in Buildings*, American Council for an Energy-Efficient Economy, Washington DC, Vol. 1, pp. 1.211-1.222.

Richalet V., Neirac F.P., Tallez F., Marco J. and Bloem J.J. 2001, Help (house energy labeling procedure) methodology and present results, *Energy and Buildings*, Vol. 33, pp. 229-233

Reddy T. A. and Claridge D. E. 1995, "Using Synthetic Data to Evaluate Multiple Regression and Principal Component Analyses for Statistical Modeling of Daily Building Energy Consumption," *Energy and Buildings*, Vol. 21, pp. 35-44.

Ruch D., Chen L., Haberl J.S. and Claridge D.E. 1993, "A Change-Point Principal Component Analysis (CP/PCA) Method for Predicting Energy Usage in Commercial Buildings: The PCA Model," *Journal of Solar Energy Engineering*, Vol. 115, pp. 77-84.

Sharp T. 1996. Energy Benchmarking in Commercial Office Buildings, Proceedings of the 1996 *ACEEE Summer Study of Energy Efficiency in Buildings*, American Council for an Energy-Efficient Economy, Washington DC, Vol. 4, pp. 321-329.

Swedish National Energy Administration 2004, *Energiläget 2003*, Stockholm (in Swedish).

Vitec AB 2004, http://www.vitec.se/enyckeln/index.htm

Vitec Fastighetssystem AB 2003, *Fastighetsklassificering e-nyckeln*, (in Swedish).

Zmeureanu R., Fazio P., DePani S. and Calla R., 1999, Development of an Energy Rating System for Existing Houses, *Energy and Buildings*, Vol. 29, pp. 107-119.