ENERGY PERFORMANCE OF BUILDINGS EVALUATED WITH MULTIVARIATE ANALYSIS

Thomas Olofsson, Jan-Ulric Sjögren and Staffan Andersson
Department of Applied Physics and Electronics
Umeå University
Sweden

ABSTRACT

Numerous problems can occur for an investigator of larger datasets, e.g. how to handle dimensionality, many variables and few observations, few variables and many observations, correlations, missing data, noise and to extract information from all data simultaneously. Multivariate analysis (MVA) is an established method for dealing with such problems.

In this work, we introduce a methodology based on MVA, which was developed to model the building energy performance from the perspective of the property holder. Data from a Swedish database of 500 buildings, which recently has been compiled and is under expansion, was used for the investigation. The available data consists of building specific information and consumption data, monitored on a monthly basis, reported by the property holder. Electrical consumption for lighting and appliances is paid by the tenants in Sweden, and is thus lacking in the database. This means that the data base just include the part of the total energy use that is paid by the property holder. With the overall goal to assess the energy use paid by the property holders, a methodology is suggested for estimating the electrical energy paid by the tenants.

At this early stage of our work, we found that the used methodology gives a fairly robust model and that the interpretation of the model is believed to be accurate in terms of comparing the energy use between different buildings.

INTRODUCTION

Investigations of the energy performance for operated buildings can be based on base lining and/or benchmarking, see for example (Fiederspiel 2002, Hicks 1999, Hinge 2004, Olofsson 2004, Sharp 1996 and Reddy 1997). With base lining we refer to an investigation where the actual consumption of a building at a specific period is compared with the expected consumption based on aggregated data from that building.

In the case of benchmarking, the comparison of the actual consumption of a building is made with the corresponding consumption of fictitious or actual operated buildings. Benchmarking data can thus be simulated and/or retrieved from actual consumption data. An investigation with simulated data has the advantage that it reveals the ideal behavior of a building with the possibility to investigate a large spectrum of possible measures. Using data from actual buildings for benchmarking has on the other hand the advantage that it reflects the influence from the actual operation of the buildings and the tenants, but with the drawback of the difficulties to identify causalities.

Analyses of the energy efficiency using monitored performance data has been investigated in earlier studies. For example, stepwise linear regression modeling was found usable to identify the strongest causes to the energy use in office buildings (Sharp 1996). This investigation is however not a simple task since such performance data are sensitive to correlations and are thus often suitable for multivariate statistical methods (Aydinalp 2002 and Olofsson 2002). Multivariate statistical methods, such as PCA has been frequently used for investigations of large data sets in natural science and engineering applications, see for example (Kettaneh 2005, Wold 2001, Apte, 1999, Olofsson 1998, Reddy 1995 and Ruch 1993). In this paper, we have used multivariate methods to develop a methodology to evaluate the building energy performance from the perspective of the property holder. The property holder is generally lacking information of domestic electrical consumption for lighting and appliances and domestic hot water preparation, paid by the tenants. In practice that has often been found treated as a constant load or based on a simple general distribution over the year.

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).

The property holder provided the building classification information and allowed the project team to retrieve monthly energy data from its building energy commissioning software. All information was entered into a SQL-database. The collected classification data included about 67 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 water consumption, electricity for building installations (fans, elevators, pumps and some lightning), energy for cooling and heating, and outdoor temperature. 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. The investigation introduced in this paper was limited to multifamily buildings, representing a little less than one third of the available buildings.

In Sweden, the electric energy supplier usually charges tenants for electric energy of appliances and lighting, not the property holder. Thus, the collected energy use data did not include the total supplied energy, but only the energy paid by the property holder, i.e. energy for heating and cooling and domestic hot water preparation.

MULTIVARIATE ANALYSIS

For describing systems in science, technology, as well as most disciplines, it can be necessary to deal with large sets of data. For analyzing larger datasets, it is generally not enough to just look at the data table. Numerous problems can occur for the investigator, e.g. how to handle dimensionality, many variables and few observations, few variables and many observations, correlations, missing data, noise and to extract information from all data simultaneously. Multivariate analysis (MVA) can be useful for dealing with such problems.

Two MVA-methods PCA and PLS, will be introduced below.

PCA:

Principal Component Analysis (PCA) can model the correlation structure of a dataset and can be used to investigate dominating variables, trends, outliers, groups, clusters and similarities as well as dissimilarities.

For a matrix where each observation is represented by one point, a mean centering procedure is used to move the coordinate system. From this procedure new vectors, referred to as principal components, can be defined. The first principal component is the line that best approximates the data. The following principal components are the orthogonal components to the previous that gives the best approximation of the data. Each principal component is associated with a loading vector, where the scores are the coordinates in the plane and the loadings defines the orientation.

PLS:

Partial least squares to Latent Structures (PLS) can be used to find relationships between two sets of multivariate data, which can be referred to as matrix **X** and matrix **Y**, and to predict one set from other new observations.

For the two matrixes, **X** and **Y**, new coordinate axes can be defined in a similar way as in PCA. The new components are referred to as PLS-components. The first PLS-component is represented by lines in X-space and in Y-space, calculated to approximate the data as well as provide a good correlation between the projections. The following PLS-components are represented as lines orthogonal to the previous that best approximate the data.

Interpretation of variable influence can e.g. be illustrated in terms of scores, VIP and regression coefficients. The score plot provides an overview of patterns between observations in the dataset. The Variable Influence on Projections, (VIP), summarizes the importance of the X-variables, and is thus a cumulative measure of the influence of a variable (Umetrics 2002). Variables with larger VIP-value than 0,8 can be regarded as significant.

MODEL

The conducted investigation was based on consumption data from 2001 and corresponding classification data for 134 buildings with a residential area larger than 75% of the rented area. Available consumption data were energy for heating and domestic hot water monitored by the property holder (E_{PH}), electricity for operating the buildings technical systems and finally the domestic cold water. In this paper, we have focused on E_{PH}.

In order to compare an individual building with others and finally to be able to pinpoint effective measures to reduce the consumption, a description of monitored data must be obtained. The description would ideally be in terms of a model where different features (actual or classification values) of the building could be used as model input. The accuracy of the model could be evaluated by simply using one part of the available data set for establishing the model and the remaining part for validation. This approach is used in PLS-software and implemented in terms of the R^2-value (goodness of fit), as a measure of the model accuracy, and the Q^2-value (goodness of prediction). The Q^2-value is obtained by cross validation. The basic idea behind cross validation is to keep a portion of the data out of the model development, develop a number of parallel models from the reduced data, predict the omitted data by the different models, and finally compare the predicted values with the actual ones.

A shortcoming with the existing database is as mentioned before, the lack of data regarding electricity used for domestic appliances and lighting. The household electricity plays an important role in the buildings energy balance since a substantial part of the household electricity finally is utilized for heating. The contribution from the gained household electricity (E_{HHE}) becomes also more important in low energy houses with a good thermal insulation. This means that E_{PH}, which is used for the entire domestic hot water preparation and only for parts of the total heating demand, is not alone a good indicator for the energy use and efficiency of a building. Thus, although E_{HHE} can be regarded as free energy from the viewpoint of the property holder, it is essential to obtain an estimate of E_{HHE} in order to be able to evaluate the buildings total energy use and efficiency.

In order to estimate E_{HHE}, we have used the following procedure. If the pattern, or monthly variation, of E_{HHE} is known and the effective overall heat loss coefficient, (K_{tot}), is assumed to be constant, the required E_{HHE} for obtaining a specific indoor temperature could be estimated together with K_{tot}. In this approach, E_{HHE} is the only free energy that is considered. In reality, solar irradiation is another source of free energy. However, for Sweden, the solar irradiation is fairly small during the winter season and becomes mainly influential during spring to autumn. This means that when analyzing data from the winter season, October to March, the influence from solar irradiation can be neglected.

For this part of the year the energy balance could thus be given as

$$\int K_{tot}(T_i - T_o)dt = E_{PH} - E_{DHW} + E_{HHE} \qquad \text{equ 1}$$

Where E_{DHW} is the energy used for domestic hot water preparation. T_i the indoor temperature and T_o the outdoor temperature.

Using data of E_{PH} from the summer months (no heating demand and only domestic hot tap water preparation) together with the domestic cold-water consumption, an estimate of the ratio of domestic hot water (m^3) to domestic cold water (m^3) was obtained. The calculations were based on an assumed domestic hot-water temperature of 55°C and the documented domestic cold-water temperature. Assuming this ratio representative for the entire year, E_{DHW} can be estimated on a monthly basis.

Since we have no data for E_{HHE} we have, in our preliminary analysis, assumed that the variation of E_{HHE} over the year is equal to that of the domestic cold water consumption for each building and that E_{HHE} for each individual building can be described by

$$E_{HHE} = S * CW \qquad \text{equ 2}$$

where S and CW is the individual scaling factor and cold water consumption in m^3 respectively, for each building. For the future, we intend to involve an E_{HHE} variation based on data from one of the large suppliers of electricity in Sweden.

Setting T_i = 22°C and assuming that K_{tot} is constant, K_{tot} and S can be estimated by a regression analysis (equ 1) on data from the period October to March. Based on the obtained value of K_{tot} we can now calculate the estimated total energy demand for heating (E_{theor}) which equals the left (or right) side of equation (1).

If the assumption behind the procedure to estimate the parameter E_{theor} is fairly accurate the obtained values of the utilized free energy (equ 2) can also be used for comparison between different buildings. A small obtained value of E_{HHE} compared to that of other buildings could indicate the presence of an excessive indoor temperature, a control system that does not work efficiently etc., since E_{HHE} is estimated under the assumption of the same constant indoor temperature for all buildings

In this early stage of our work, the first aim was to develop a model for E_{PH} with a high R^2-and Q^2-values. The model is used for comparison, but ideally also for predicting the contributions from different model parameters. Based on the available data of building specific parameters and the calculated E_{theor} an PLS-model was developed for E_{PH} per square meter and year. The use of the calculated E_{theor} improved the model and its Q^2-value considerably.

Today, in most assessments of the performance of individual buildings in Sweden, the contribution from the electrical energy paid by the tenants to the total energy demand, is either assumed constant or not included.

Multivariate model

The conducted multivariate analysis was based on an PLS on two blocks, the first where the **X**-matrix with the documented building specific data (see below) as well as E_{theor}. The second, the **Y**-matrix includes one vector with corresponding E_{PH}. The PLS-analysis was conducted with the commercial software Simca 10.0 (Umetrics 2002) on data of buildings with a residential area larger than 75%.

Below we introduce the treatment of the modeled parameters used in the PLS-model.

Quantitative parameters:

Y: Yearly data for E_{PH}. The main supply sources were district heating, electricity, gas, and oil.
X1: Estimated yearly data for E_{theor}, according to the procedure introduced above.
X2: Year of construction

Qualitative parameters:

X3: Thermal inertia
M - Medium
H – High
X4: HVAC-system
N - Natural ventilation
E - Exhaust fan
NE- Combined Natural ventilation and Exhaust fan
SE - Supply and exhaust fan
SERC - Supply and exhaust fan with Recuperative heat exchanger
SERG - Supply and exhaust fan with Regenerating heat exchanger
X5: Electrical heaters
Y-Yes
N-No
X6: Electrical boiler
Y-Yes
N-No
X7: Gas boiler
Y-Yes
N-No
X8: District heating system
Y-Yes
N-No
X9: Heat pump system
Y-Yes
N-No
X10 Building operation organization
I – Internal personal
E- External personal

RESULTS AND DISCUSSION

In order to get a comparative set of objects for the investigation, buildings with less than 75% residential rented areas were excluded. Additionally, buildings without continue documented monthly consumption data for 2001 were also excluded. Finally, 134 objects remained.

The investigated variables were selected from a preliminary analysis, when variables that were not found to have a significant improvement of Q^2 or did not included unique information were excluded. Finally, ten variables remained, introduced above. A more thoroughly investigation on the available dataset will be conducted in a future step of the project.

A PCA was conducted on the **X** –matrix (the parameters X1 to X10), to get a general overview of groupings and eventual outliers. A score plot of the first two principal components, PC1 and PC2, indicates that there are no obvious groupings, see figure 1. The score plot also indicates that the data are rather well collected.

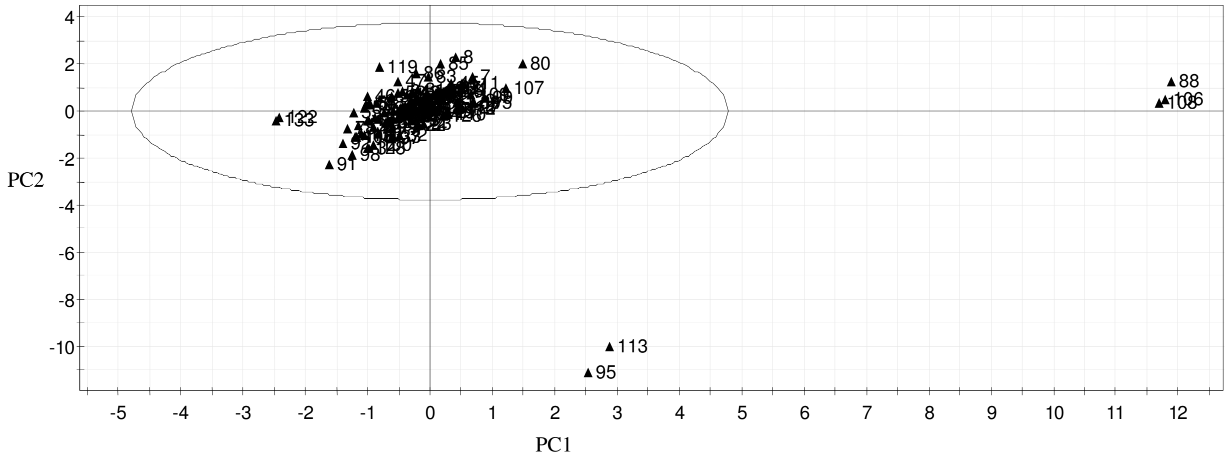

Figure 1. A plot of the scores of two first Principal Components PC1 and PC2. The buildings are represented by no 1 to 134 and in general found well collected inside Hotelling's T^2 .

Based on the score plot in figure 1, a few buildings was identified, outside Hotelling's T^2 and accordingly referred to as outliers. The same pattern was also indicated from a companion score plot of the combined **X**-matrix and **Y**-matrix, see figure 2, where the **Y**-matrix, is the vector with E_{PH}. The outliers were found, in most cases, to be described with extraordinary size, age, etc. Thus, the outliers were not excluded from this analysis. In the future, when more buildings have been registered in the database, improved interpretation can be expected.

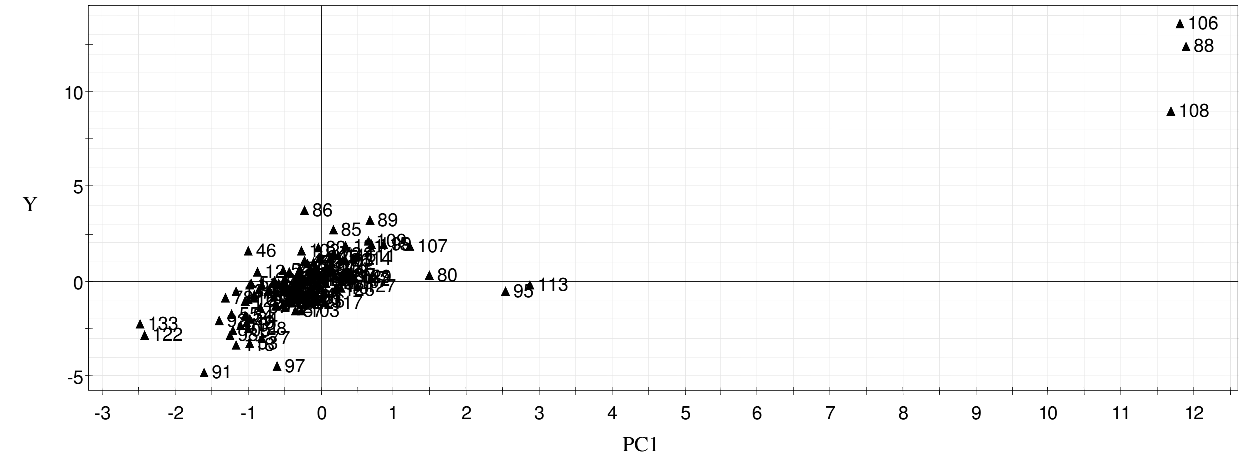

*Figure 2. A score plot on the first principal components of the dataset in the **X**-matrix, in terms of PC1 and **Y**-matrix, in terms of **Y**. The buildings are represented by no 1 to 134 and found well collected.*

A PLS was conducted on the **X**-matrix and **Y**-matrix yielding an R^2=0,808 and Q^2=0,750, based on two components. The R^2 value of the PLS-model was found satisfactory and the fact that Q^2 is fairly equal to R^2 indicate a fairly robust model.

In order to evaluate the variable influence an VIP for the investigated data is shown in figure 3. Larger VIP than 0,8 can be assigned as significant, The most significant variable is the estimated E_{theor}, X1. The second most significant variable is the district heating, X7. The third most significant variable is gas heating, X8. For this investigation, the remaining investigated variables are not found significant, although they improved R^2 and Q^2.

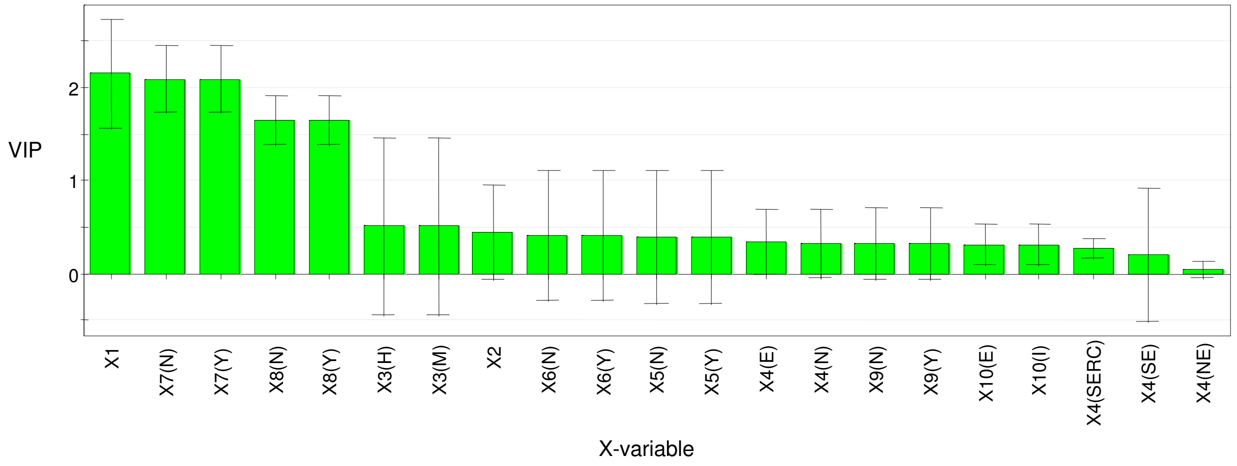

Figure 3. The variable importance, VIP, conducted in the investigated dataset.

The estimated regressions coefficients are illustrated in figure 4. The bars indicate the scaled and centered size and sign of the variable loadings, to model and predict the energy consumption, in terms of E_{PH}. For this particular investigation it is shown that estimated E_{theor}, X1, increases E_{PH}, in accordance with the definition, the district heating, X7(Y), increases E_{PH} and the gas heating, X8(Y) decreases E_{PH}. The latter behavior could be a true effect or due to the used conversion factor from consumed gas to energy.

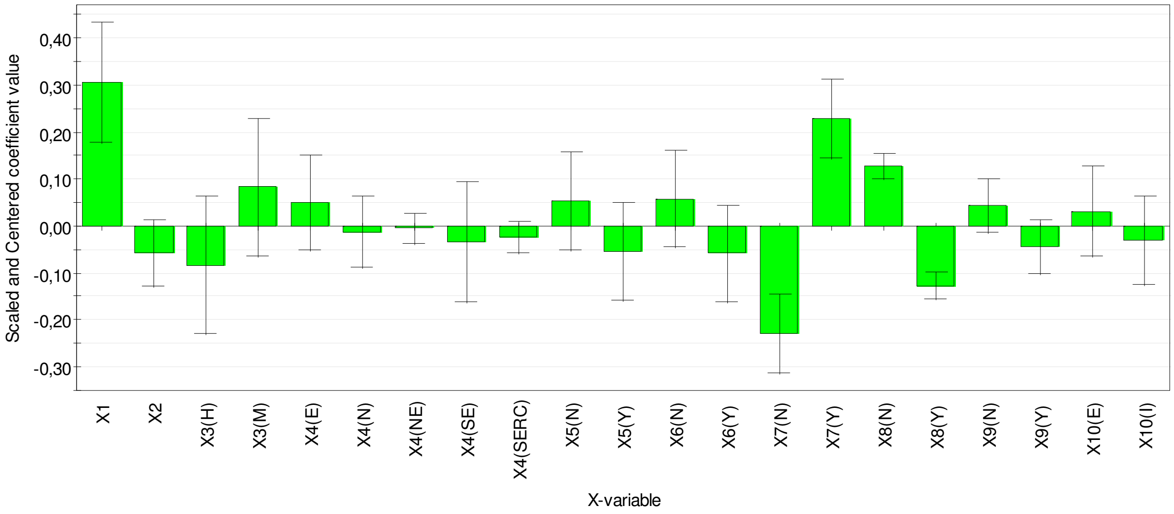

Figure 4. Scaled and centered coefficient value of the X-variables.

In figure 5, the residual, as the deviation between the model and the actual E_{PH}, is shown for the 134 buildings in a histogram.

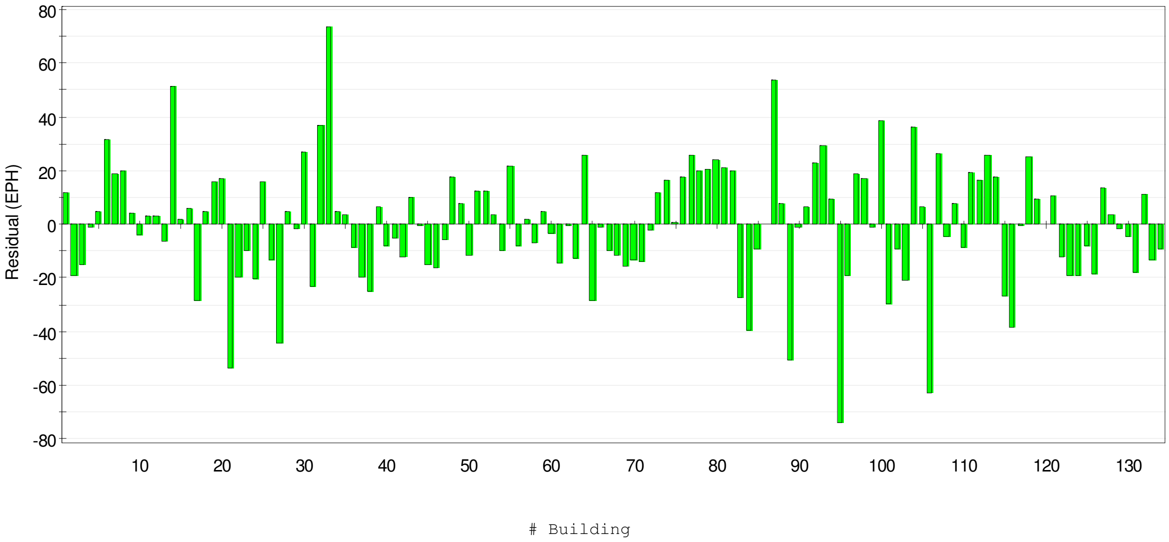

Figure 5. The size and sign of the residual, in terms of model and the actual E_{PH}, for the 134 investigated buildings.

Since we, as mentioned before, lack information about E_{HHE} the deviation between model and prediction could be due to a smaller or larger E_{HHE} than what is normal for the investigated buildings or due to varying indoor temperatures and/or other missing variables. The residuals shows, as seen in figure 6, a fairly normal distribution, which could be taken as an indication that the missing inputs also exhibits a normal distribution, a distribution which E_{HHE} and T_i is expected to exhibit.

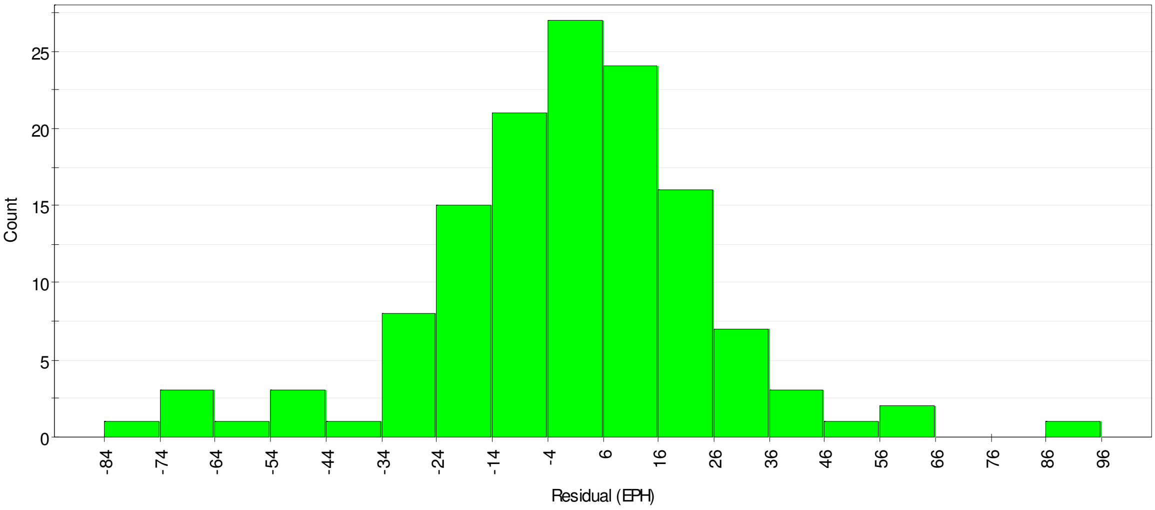

Figure 6. Histogram for the variation of the residual.

From this line of reasoning the residuals could more or less reflect the differences in E_{HHE} and T_i which for most building are ±35 kWh/m^2,year, which is not unreasonable. The regression coefficients received from the PLS offers of course a possibility to predict how the energy consumption will change due to different measures, but before this is done a more thorough investigation has to be performed.

CONCLUSION

In this paper we have described the first steps taken, based on an MVA-method, to deal with benchmarking of residential buildings where only parts of the total energy use is known. For the investigated multifamily residential buildings, the tenants paid the bills for the household electricity and therefore information of this supplied energy is lacking, which is the normal situation in Swedish residential buildings.

From the available data of the 134 buildings and assumptions of a constant overall heat loss coefficient, a constant domestic hot-water to cold-water ratio and a known household energy pattern a theoretical estimate of the total energy use was made. This theoretical estimate of the total energy demand of the buildings was then used as an input together with other classification data to improve the PLS- model for the yearly energy use per square meter monitored by property holders, E_{PH}. The obtained PLS-model had fairly good and equal Q^2- and R^2-values, 0.750 and 0.808 respectively, which indicates a fairly robust model.

In addition, the distribution of the deviation between actual and modeled E_{PH} have a fairly normal distribution, which could be expected since the lack of two important inputs, household electricity and indoor temperature is expected to be normally distributed.

At this stage of our work the interpretation of the model is only believed to be accurate in terms of comparing the energy monitored by different property holders. For the next step, to be able to use the model for predicting the effect from different measures additional, work has to be done.

ACKNOWLEDGMENT

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), The Development Fund of the Swedish Construction Industry (SBUF) and NCC AB for their financial support of this work.

REFERENCES

Apte M.G and Daisey J.M. (1999) VOCs and 'Sick Builing Syndrome': Application of a New Statistical Approach for SBS Research to U.S. EPA BASE Study Data, report LBNL-42698, Lawrence Berkeley National Laboratory.

Aydinalp M., V. Ugursal i. and Fung A.S. (2002), Modeling of the appliance, lighting, and space-cooling energy consumptions in the residential sector using neural networks, Applied energy, Vol. 71, Pages 87-110.

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 von Neida B. (1999). An Evaluation of Americas First Energy Star ® Buildings: The Class of 1999, Commercial Buildings, Program Design, Implementation and Evaluation, pp. 4.177-4.185.

Hinge A. (2004), Comparing Commercial Building Energy Use Around the World, Proceedings of the 2004 ACEEE Summer Study of Energy Efficiency in Buildings. American Council for an Energy-Efficient Economy, Washington DC, pp. 4.136-4.147.

Kettaneh N., Berglund A. and Wold S. (2005), PCA and PLS with very large data sets, Computational statistics and Data Analysis, Vol. 48, pp. 69-85.

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. 211-222.

Olofsson T. and Andersson S. (2002), Overall Heat Loss Coefficient and Domestic Energy Gain Factor for Single-Family Buildings, Building and Environment Vol. 37 (11), pp. 1019-1026.

Olofsson T., Andersson S. and Sjögren J.-U. (2004), Multivariate Methods for Evaluating Building Energy Efficiency, Proceedings of the 2004 ACEEE Summer Study of Energy Efficiency in Buildings. American Council for an Energy-Efficient Economy, Washington DC, pp. 4.265-4.274.

Olofsson T., Meier A. and Lamberts R. (2004), Rating the Energy Performance of Buildings, The International Journal of Low Energy and Sustainable Buildings, Vol. 3.

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.

Reddy T. A. and D. E. Claridge (1994), 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.

Reddy T.A., Saman N.F., Claridge, Haberl J.S. Turner W.D. and Chalifoux A.T. (1997), Baselining Methodology for Facility-Level Monthly Energy Use – Part 1: Theoretical Aspects, ASHARAE Transactions BN-97-16-4 (4089), pp. 1-12.

Ruch D., Lu Chen, J. S. Haberl and D. E. Claridge (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.

Umetrics AB, (2002), User's Guide to SIMCA-P, SIMCA-P+, Version 10.0, Umeå, Sweden.

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

Vitec AB (2005), http://www.vitec.se/enyckeln/index.htm

Wold S., Sjöström M and Eriksson L (2001), PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems, Vol. 58 pp. 109-130.