Statistical Modeling of Daily Energy Consumption in Commercial Buildings Using Multiple Regression and Principal Component Analysis

Jianxun Wu, T. Agami Reddy, and David Claridge

Energy Systems Laboratory, Mechanical Engineering Department
Texas A&M University, College Station, TX

ABSTRACT

Statistical models of energy use in commercial buildings are being increasingly used not only for predicting retrofit savings but also for identifying improper operation of HVAC systems. The conventional approach involves using multiple regression analysis to identify these models. However, such models tend to suffer from physically unreasonable regression coefficients and instability due to the fact that the predictor variables (i.e., climatic parameters, building internal loads, etc.) are intercorrelated. A relatively new approach proposed to circumvent these drawbacks is principal component analysis. The objective of this paper is to evaluate the multivariate regression and the principal component analysis approaches, using measured whole-building energy use data from a large commercial building in central Texas. For the types of correlation strengths among the regressor variables present in our data, we find that there does not seem to be much justification in selecting the principal component analysis approach. A more careful and elaborate investigation using data sets which exhibit a wide range of multicollinearity strengths is required in order to ascertain when principal component analysis yields predictive models superior to those of a multiple regression approach.

INTRODUCTION

The statistical modeling approach for predicting energy consumption in commercial buildings has been studied and used in recent years both for estimating retrofit savings and for identifying improper HVAC operation (e.g., Haberl and Claridge, 1987, Claridge et al., 1990, Kissock et al., 1992). Conventionally, simple or multiple linear regression analysis (MRA)* is applied to the data set to identify a model based on the least-squares principle. The MRA method is relatively simple to understand, easy to implement and quite effective for many purposes. However, with intercorrelated regressor parameters, model stability and positive identification of the importance of individual predictors become uncertain. A technique in statistics which has the potential to overcome these difficulties is Principal Component Analysis (PCA). Only recently has PCA been applied to whole-building electricity consumption for a grocery store (Ruch et al., 1990).

* Many commercial buildings exhibit change point behavior, i.e., a segmented linear model is more appropriate than a linear model (Ruch and Claridge, 1991). We shall, however, not address such models in this paper given the preliminary nature of this study.

The main objective of this study is to evaluate and assess the scope and benefits of adopting PCA vs. MRA using measured data from a large commercial building in central Texas. Other issues concerning the validity of linear models of whole-building energy use in commercial buildings are explored with data generated synthetically by a detailed simulation code.

MATHEMATICAL BACKGROUND

In this section, the mathematical fundamentals of both the MRA and PCA approaches will be briefly described.

1. MRA Approach

The regular linear MRA model can be expressed as (Draper and Smith, 1981):

$$y = a_0+a_1*x_1+a_2*x_2+\ldots+a_n*x_n \qquad (1)$$

where

y - dependent or response variable, (e.g., whole-building electricity use, or hot water consumption or chilled water consumption),

Proceedings of the Eighth Symposium on Improving Building Systems in Hot and Humid Climates, Dallas, TX, May 13-14, 1992

$x_1, x_2, \ldots, x_n$ - regressor variables (e.g., ambient dry-bulb temperature, humidity, solar radiation, internal loads of the building)

$a_0, a_1, \ldots a_n$ - regression coefficients.

The following statistical indices are often used to evaluate an MRA model (Draper and Smith, 1981):

(1) Coefficient of determination of the model (R square) which is a measure of goodness-of-fit of the model to the data;

(2) Root Mean Square Error (RMSE) which is a measure of the mean difference between the model and the data;

(3) Coefficient of Variance (CV) which is the normalized RMSE i.e., the RMSE divided by the mean value of the dependent variable;

(4) Standard error.

MRA is a standard feature in many computer packages (for example, SAS 1989) where the above indices and others are calculated. Basic MRA without all the detailed error diagnostics and statistical indices (except R square) can even be done with most spreadsheet programs using micro-computers.

Classical regression analysis assumes the regressor variables to be independent of one another. However, this is not the case in many physical problems. Multicollinearity between response variables results in large uncertainty bounds for the regression coefficients and also in model uncertainty - which is especially crucial when a model identified from a certain data set is used to predict future values.

2. PCA Approach

The PCA method is a classical multi-variate technique which originated with Pearson in 1901 as a means of fitting planes by orthogonal least squares. It was later used by Hotelling in 1933 for the purpose of analyzing covariance and correlation structuring. Since then it has become increasingly popular in multi-variate statistical theory and can be used to overcome multicollinearity effects. In essence, PCA (see any appropriate textbook, for example Jolliffer, 1986; Van Rijckevorsel and de Leeuw, 1988; Flury, 1988; Jackson, 1991; Daultrey, 1976) is a statistical technique useful for describing and summarizing data. It takes a group of "n" variables and re-expresses them as another set of "n" indices, each of which represents a linear combination of the original variables. These indices, known as principle components (PCs) have several useful properties; they are uncorrelated with one another and they are ordered so that the first PC explains the largest proportion of the variation of the original data, the second PC explains the next largest proportion and so on. When the original variables are highly correlated, the variance of many of the later PCs will be so small that they can be ignored. In our particular problem, the advantage of resorting to PCA is not so much in its ability to summarize data but rather in being able to remove the multicollinearity effects in the regressor variables (via the PCs) and order them.

The technique adopted by Ruch et. al (1990) for modeling daily whole-building electricity use data was (1) to identify the most influential PCs as explained above, (2) perform multiple regression of electricity use versus these influential PCs, and (3) transform the regression coefficients and the influential PCs back in terms of the physical regressor variables. Details of this technique applied to energy data from buildings, including a detailed numerical example, are given in Chen (1991). It is obvious that the PCA approach is more demanding in time, effort and statistical understanding than the MRA technique. Whether these can be justified by increased robustness of the models identified for applications involving building energy use remains uncertain--an issue which this study strives to address.

The mathematical treatment of PCA is based on characteristic roots and vectors of positive definite symmetric matrices. PCA regression involves the following statistical measures: (a) simple statistics for the parameters in the models: mean values and standard deviation; (b) a correlation matrix for all parameters in the models; (c) eigenvalues of the correlation matrix for every principal component: eigenvalues, their differences, proportions and cumulatives; (d) eigenvectors showing the relationship between principal components and every parameter in the model; (e) analysis of variance, R square, adjusted R square, root mean square error RMSE, coefficient of variance CV; and, (f) parameter estimates and probabilities for every parameter in the models and for every principal component.

DESCRIPTION OF THE ENGINEERING CENTER

The Engineering Center is a 324,400 square foot (30,138 m^2) building which houses offices, classrooms, laboratories and a large central computer facility. It was built in early the 1970s and is located on the Texas A&M University campus. It is a 4-story rectangular structure (plus an unconditioned basement floor) with the long axis along the N-E to S-W direction. One of the distinguishing features of the building is a large centralized, three-story atrium that provides access to the

Proceedings of the Eighth Symposium on Improving Building Systems in Hot and Humid Climates, Dallas, TX, May 13-14, 1992

surrounding classrooms and offices. Approximately 10% of the surface is glazed, so solar radiation is not a major source of heat load.

The Engineering Center classrooms and labs are scheduled from 7:30 a.m. to 6:30 p.m. weekdays, but labs are in use continuously. Chilled water, hot water and electricity are provided from the physical plant via an underground tunnel.

There were 12 identical CAV systems with 40 HP fans rated at 35,000 cfm and 8 smaller air handlers (27 HP average) located around the perimeter of the building during the time that the data analyzed in this paper were taken. The air intakes provide about 10% outdoor air when fully open.

The Engineering Center has been extensively retrofitted as part of the LoanSTAR program (Verdict et al., 1990). Whole-building energy use as well as sub-metered data and several climatic parameters are monitored hourly. The energy use and the system behavior of this building have been extensively reported elsewhere (Bronson et al., 1992; Katipamula and Claridge, 1992). In this study, we shall limit our investigation to daily data from the period, September 1988 to February 1990 - a period of about six months for which clean data are available.

IDENTIFICATION OF IMPORTANT REGRESSOR VARIABLES

There are several factors which affect the air-conditioning energy consumption of commercial buildings. Essentially, they can be divided into two kind of parameters: weather dependent and weather independent.

Primary weather dependent factors include:(1) ambient temperature T; (2) relative humidity RH (alternatively, the wet-bulb temperature or the specific humidity SPH); (3) air pressure P; (4) enthalpy ENTH which can be computed if T and RH are known (Zaikong et al., 1985); (5) solar radiation SOL; and, (6) wind speed WIND.

Weather independent variables include: (1) time schedules: regular (weekdays, weekends, holidays) and stochastic; (2) lighting loads; (3) occupant loads: sensible and latent; (4) internal equipment and appliances, and others. The sum of lighting loads and internal equipment will be called lights and receptacle load LR.

Weather dependent parameters are stochastic, i.e., they vary randomly from time to time, and cold/hot weather fronts strongly affect the regular pattern of the building energy consumption models.

It is statistically unsound and also unnecessary to include all the above parameters in regression models for building energy consumption. There are a number of ways of identifying the parsimonious set of regressor variables during multiple regression. The most commonly used technique is to use step-wise regression wherein the computer code itself decides whether to include a variable or not based on an F-Test (Draper and Smith, 1981). This technique may pick different sets of variables when different data sets are used. In order to overcome this limitation, selection of important regressor variables in this study has been based simply on the strength of the correlation coefficients.

Table 1 lists the correlation coefficients of the regressor variables enumerated earlier versus chilled water energy use, CW (MBtu/h) and hot water energy use, HW (MBtu/h). We note that clearly the ambient temperature T (°F) is the most important parameter, followed by specific humidity SPH (lb./lb.), solar radiation SOL (W/square ft) and internal load LR (kWh).

Table 1. Absolute values of the correlation coefficients of various regressor variables versus chilled water and hot water energy use.

Parameter	CW	HW
T	0.905	0.919
SPH	0.788	0.690
SOL	0.321	0.443
RH	0.304	0.213
WIND	0.066	0.170
LR	0.368	0.404

Time plots of CW and T, and HW and T are shown in Figures 1 and 2 respectively.

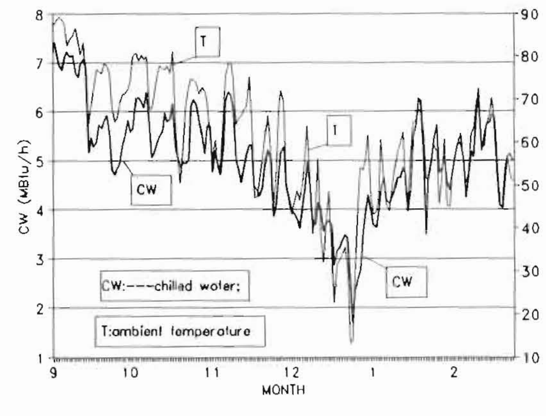

Figure 1. Time plots of daily chilled water use (CW) and ambient temperature (T).

Proceedings of the Eighth Symposium on Improving Building Systems in Hot and Humid Climates, Dallas, TX, May 13-14, 1992

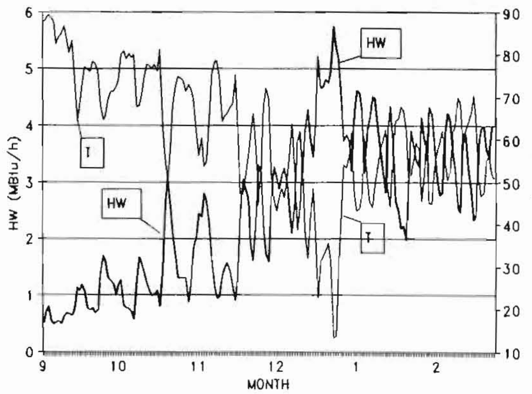

Figure 2. Time plots of daily hot water use (HW) and ambient temperature (T).

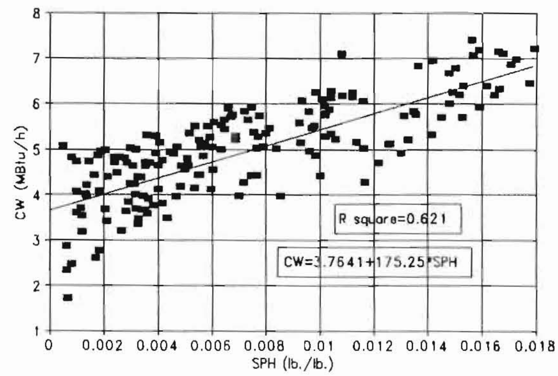

Figure 5. Scatter plot of daily chilled water use (CW) versus specific humidity (SPH).

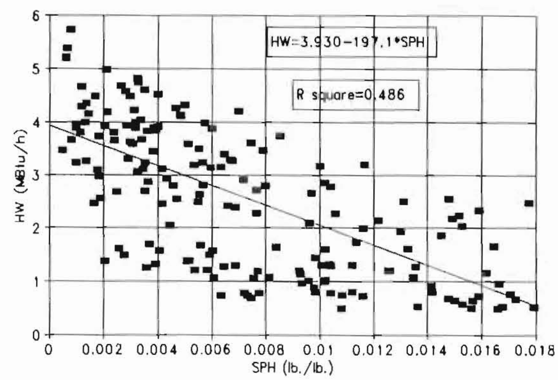

Figure 6. Scatter plot of daily hot water use (HW) versus specific humidity (SPH).

Scatter plots of CW and HW versus T and SPH, shown in Figure 3-6, illustrate the relationship between these variables.

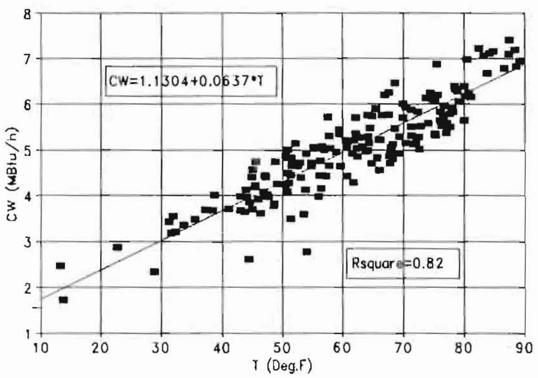

Figure 3. Scatter plot of daily chilled water use (CW) versus ambient temperature (T).

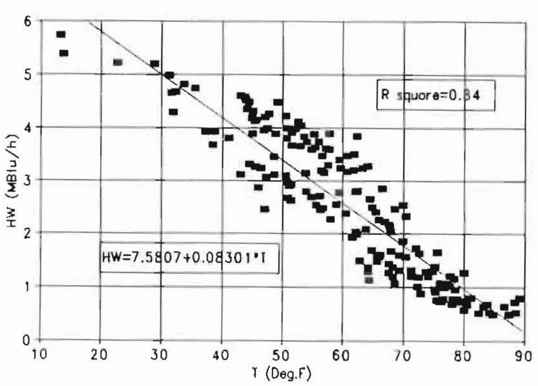

Figure 4. Scatter plot of daily hot water use (HW) versus ambient temperature (T).

We note that CW use increases with both T and SPH while that of HW decreases with both T and SPH. This is consistent both with past experience on regressing building energy use (for example, Fels, 1986; Kissock et al., 1992) and with engineering principles based on our physical understanding of the operation of HVAC systems in buildings. Thus in the analysis that follows, we shall limit our regression models for CW and HW to only the following four regressor variables: T, SPH, SOL and LR.

VALIDITY OF LINEAR MODELS

A fundamental premise in the development of statistical models for CW and HW is that these quantities are linearly affected by the regressor variables selected. Other than the simplest case of energy consumption in residences (Fels, 1986) where concurrent heating and cooling does not occur as it does in commercial buildings, it is difficult to prove this linearity from physical or engineering considerations. Most studies to date presume

Proceedings of the Eighth Symposium on Improving Building Systems in Hot and Humid Climates, Dallas, TX, May 13-14, 1992

such a relationship, a concept which has gained credibility due to the fact that most energy use models for commercial buildings have high R square values when linear regression is applied to the measured data (Kissock et al., 1992). In this subsection we shall briefly describe and illustrate a way by which such linearity can be investigated.

The conventional way of studying system behavior or response to certain forcing functions is to perform controlled experiments where one parameter is varied while keeping other parameters constant. From this, the effect of that particular parameter on the system behavior can be deduced. Such controlled experiments have been done in residences with good results (e.g., Sonderegger, 1978 and Subbarao, 1988). However, it is difficult if not impossible, to perform such experiments on large commercial buildings. Use of synthetic data looks like the most promising approach to this problem.

Synthetic data is generated using a computer, not obtained from experimental measurements. We have chosen a detailed simulation code, namely DOE-2 (LBL, 1981) which permits hourly energy consumption values of commercial buildings to be predicted from a physical description of the building and of the HVAC system, from climatic parameters and from schedules of internal loads and building operation. Hourly simulation codes are effective tools for design purposes, and most recently, have been used to evaluate retrofits in existing buildings as well (Bronson et al., 1992). However, for the evaluation to be meaningful, agreement between measured and simulated data must be verified, for which an iterative process called "calibration" of the input data to the simulated code is required. Such a calibration run (called Run 1) was performed for the Engineering Center using climatic and energy use data from September, 1989 to February, 1990 by Bronson et al. (1992). We used the same calibrated input deck as the baseline simulation for this study. Figures 7 and 8 illustrate how the simulation values of daily CW and HW compare with the measured values. We note very good agreement between simulated and measured values, a fact also supported by the high R square values between both these sets of values.

Once the input parameters of the simulation code have been calibrated to represent realistic building and HVAC system behavior, the effect of specific parameters on CW and HW use can be studied. For example, another run (say Run 2) can be performed with the specific parameter removed (i.e., the system is no longer subject to that particular forcing function) with all other parameters left unaltered. A linear relationship between the "extra" energy consumption, i.e., the difference in energy use of Run 1 and Run 2, and the specific parameter would be an indirect justification of selecting a linear regression model between energy use and that particular parameter.

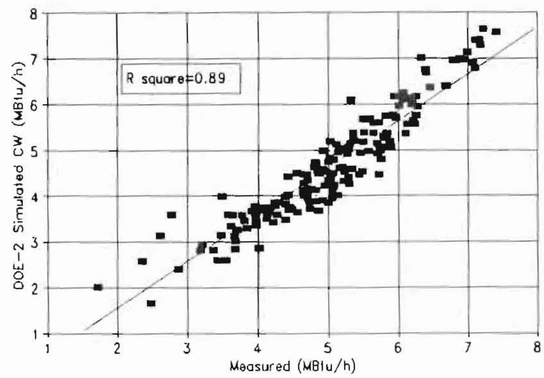

Figure 7. Cross plots of daily chilled water use illustrating how well synthetic data generated by DOE-2 predicts measured use

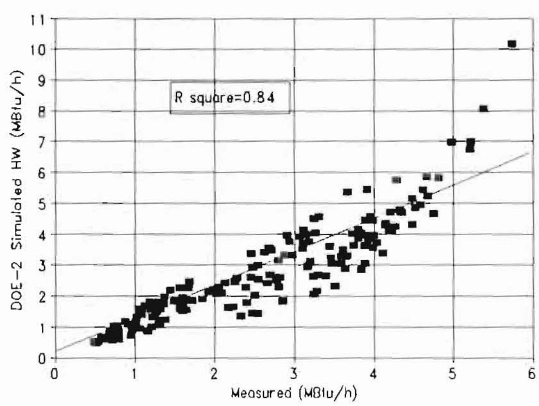

Figure 8. Cross plots of daily hot water use illustrating how well the synthetic data generated by DOE-2 predicts measured use

A procedure as described above was performed with our calibrated DOE-2 simulation code for the Engineering Center with the entire solar radiation SOL data set to zero. How the simulated values of (Run 1 - Run 2) for CW and HW vary with solar radiation are shown in Figures 9 and 10. We detect strong linearity between the "extra" CW and SOL (R square = 0.724) while that between the "extra" HW and SOL is less strong, but still distinct (R square = 0.405).

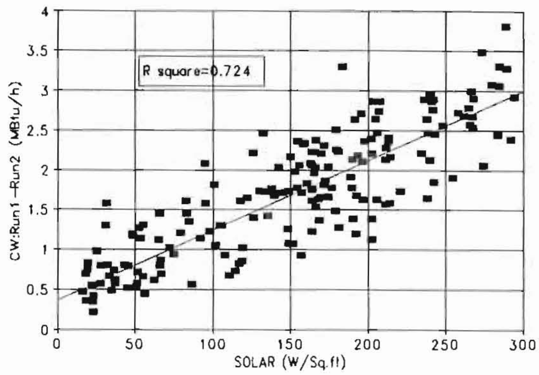

Figure 9. Scatter plot of the "extra" daily chilled water use versus daily solar radiation using synthetic data. A fairly strong linear correlation can be seen.

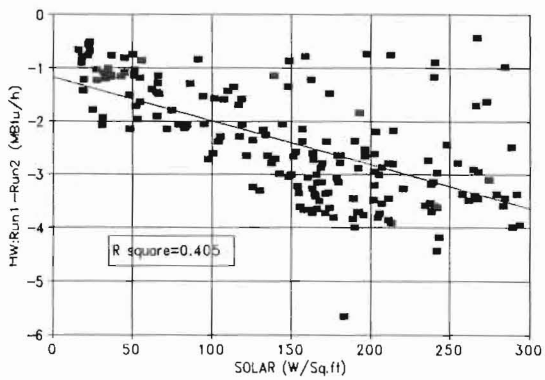

Figure 10. Scatter plot of the "extra" daily hot water use versus daily solar radiation using synthetic data. A less obvious but nevertheless distinct linear correlation can be detected.

The above procedure of verifying the validity of using linear models for CW and HW use in commercial buildings needs to be extended to include the effect of parameters other than SOL, something which is less straight-forward. This approach is currently being investigated and is presented here more as a conceptual approach for ascertaining the validity of linear functions for the statistical models.

EVALUATION OF THE STATISTICAL APPROACHES

The final objective of the statistical regression approach is a sound predictive model for daily whole-building energy use. Therefore, our model must first provide a good fit to the current data, and secondly, be a reliable predictor of future consumption, provided of course that the operation and scheduling of the building are unaltered. How the MRA and PCA approaches compare with each other in both these aspects will be investigated in the following section.

Before proceeding to do so, we should, however, ascertain the extent to which the selected regression variable in our data set are intercorrelated. If no collinearity exists, a meaningful evaluation of MRA and PCA approach would not be possible. Table 2 presents the correlation coefficients between the four variables T, SPH, SOL, and LR for the entire six months of data. We note that T and SPH are strongly correlated, that T and SOL are moderately so, and that the others show little or no correlation. A rule of thumb (Draper and Smith, 1981) states that multicollinearity is likely to be a problem if the simple correlation between two variables is larger than the correlation of one or either variable with the dependent variable. By this token, collinearity between say, T and SPH, and T and LR, may not be a problem if MRA is used. (See Tables 1 and 2).

Table 2. Correlation coefficients between the regressor variables using six months daily data.

	T	SOL	SPH	LR
T	1.00	0.423	0.737	0.245
SOL		1.00	-0.107	0.084
SPH			1.00	0.150
LR				1.00

The entire data set of daily measurements for the Engineering Center during the period from September, 1989 to February, 1990 was used to perform an MRA. The regression coefficients and the corresponding standard errors are given in Table 3 while the pertinent statistical indices are shown in Table 4. We note that the R square values are close to 0.9, indicating that the models fit the data very well. We also note that all the regression coefficients are physically consistent in sign and that the standard errors, except for SPH in the CW model, are generally low.

Table 3. Regression coefficients and their standard errors for all six months of data using MRA. Standard errors are expressed as fractional values of the corresponding coefficient.

	CW		HW	
Variable	Coefficient	Std. Error	Coefficient	Std. Error
Intercept	16.262	0.33	212.765	0.03
T	0.9900	0.11	-1.5660	0.08
SPH	1749.06	0.51	-931.48	0.39
SOL	0.02780	0.18	-0.0448	0.37
LE	0.00173	0.18	0.00276	0.13

Table 4. Statistical indices of regression models using the entire six months of daily data.

	MRA		PCA (3 PCs)	
	CW	HW	CW	HW
Adj. - R^2	0.855	0.876	0.853	0.864
RSME (MBtu/d)	9.59	11.31	9.65	11.86
CV (%)	7.98	18.58	8.03	19.50
RE (%)	0.704	4.795	0.800	4.237

One of the important purposes of using PCA and model optimization is to arrive at the minimum number of principal components (PCs) which satisfactorily account for or explain the variance in the data set of regressor values. A PC with a sufficiently low variance rank can be eliminated without undue loss of information. This elimination will result in increased model stability, albeit at the expense of a slight reduction in goodness-of-fit to the measured data.

Table 5 presents the eigenvalues of the four PCs along with their rank. We note that three PCs explain 97% of the variance and so retaining these three PCs may be a logical choice.

Table 5. Eigenvalues of the principal components.

	Eigenvalues	Proportion	Cumulative
PC1	1.904	0.476	0.476
PC2	1.093	0.273	0.749
PC3	0.904	0.226	0.975
PC4	0.099	0.025	1.000

From Table 6, which lists the eigenvectors of all four PCs, we note that the second and third PCs are essentially the SOL and LR parameters respectively. PC1 has a more equitable weight, though T and SPH are definitely more important.

Table 6. Eigenvectors of the principal components.

	PC1	PC2	PC3	PC4
T	0.694	0.046	-0.194	-0.692
SOL	0.284	0.857	-0.178	0.392
SPH	0.584	-0.509	-0.187	0.604
LR	0.311	0.070	0.946	0.051

Another criteria for choosing the relevant PCs is to look at how the model R square varies when the PCs are successively dropped from the model. From Table 7 which presents the results of such an approach, we note that dropping PC3 may be another choice. How the regression model R square varies when a different number of PCs are chosen is shown in Table 7.

Table 7. Variation of model goodness-of-fit i.e., adjusted R square values with variable number of PCs.

	CW	HW
PC1	0.850	0.860
PC1, PC2	0.853	0.864
PC1, PC2, PC3	0.854	0.864
All 4 PCs	0.855	0.876

We observe a slight decrease in the R square values as the number of PCs in the model is decreased. Even with only one PC, the R square values have decreased by less than 2 percentage points. As born out by statistical theory, we note that with all 4 PCs present in the model, the R square values of CW and HW models are identical to those of the MRA models.

The regression coefficients and the corresponding standard errors of the PCA using all 4 PCs are shown in Table 8. Because the PCs are uncorrelated, the PCs with little predictive power and low variance rank can be dropped from the regression equation without changing the regression coefficients of the remaining PCs. Hence, these coefficients are unaltered if either PC3 or PC4 is omitted.

We note that the standard errors of the first three PCs are lower than those of MRA (see Table 3) while PC3 has very large standard errors. A logical choice would thus be to drop PC3 (though based on the rank variance criteria, PC4 would be the PC to discard from the model).

Table 8. Regression coefficients and their standard errors for all six months of data using PCA with all four PCs present. Standard errors are expressed as fractional values of the corresponding coefficient.

	CW		HW	
Variable	Coefficient	Std. Error	Coefficient	Std. Error
Intercept	120.22	0.00	60.842	0.01
PC1	16.8536	0.03	-21.617	0.03
PC2	-1.4272	0.48	-2.227	0.36
PC3	-0.9206	0.82	-0.2211	4.02
PC4	-4.052	0.56	11.639	0.23

From Table 4, which enables comparison of the statistical indices describing the MRA and PCA modeling approaches, we note that PCA has sacrificed a little in terms of goodness-of-fit hoping that this will lead to enhanced model stability. In order to evaluate the latter, we have decided to proceed as follows:

(a) choose only the first two months of daily data and identify appropriate CW and HW models following the MRA and PCA approaches as described above;

(b) use these models to predict daily energy consumption data over the next four months;
(c) compute and compare the RMSE and CV values of both these modeling approaches.

Table 9. Comparison of the MRA and PCA approaches as predictive tools. Regression models were identified from two months of daily energy data and were used to compare daily energy consumption predictions over the next four months for which measured data were available.

	RMSE (MBtu/d)		CV (%)	
	CW	HW	CW	HW
MRA	11.86	23.84	10.81	30.72
PCA (with 3 PCs)	12.55	25.37	11.43	32.69

Table 9 presents the RMSE and CV values for the MRA and PCA approaches (using 3 PCs only) during the four months from November, 1989 to February, 1990.
We note that, surprisingly, MRA is still superior to the PCA despite multicollinearity effects being present in our regressor data set (the correlation coefficients for the two-month data set are akin to those for the entire data set which are shown in Table 2). There does not seem to be any justification in selecting PCA over MRA. This rather negative result highlights the fact that collinearity among variables may not be an important issue unless the strengths of the correlation are relatively high. A more careful and elaborate investigation using data sets which exhibit a wide range of correlation strengths among the regressor variable is required in order to ascertain the exact magnitude of these correlation coefficients beyond which better predictive models can be identified using PCA rather than MRA.

ACCURACY OF "MRA" MODEL PREDICTION

In this section, we shall illustrate how model prediction accuracy varies from month to month. An MRA model was identified from one month's daily data and then used to predict daily values over the remaining five months. How well the model-predicted values compare with actual measured values has been quantified in terms of the CV computed on a month-by-month basis. This procedure has been repeated five times, choosing a different monthly set to identify the MRA model.

The CV values thus obtained for daily chilled water energy use are given in Table 10 and also shown in Figure 11. As expected, the CV values are lowest for the month used to identify the model (these values are shown underlined in Table 10).

Table 10 Predictive accuracy expressed as CV values (%) of a model identified by MRA using one month's data.

	Month used for Prediction					
Month used for model identification	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.
Sept.	<u>2.35</u>	6.22	10.61	15.98	15.80	14.57
Oct.	6.54	<u>3.35</u>	8.23	17.90	7.62	7.34
Nov.	11.42	5.71	<u>5.33</u>	9.94	10.94	11.02
Dec.	18.07	13.11	8.67	<u>5.74</u>	16.73	17.94
Jan.	8.10	4.42	11.29	22.91	<u>6.96</u>	6.06
Feb.	8.08	8.86	11.85	20.37	8.45	<u>4.22</u>

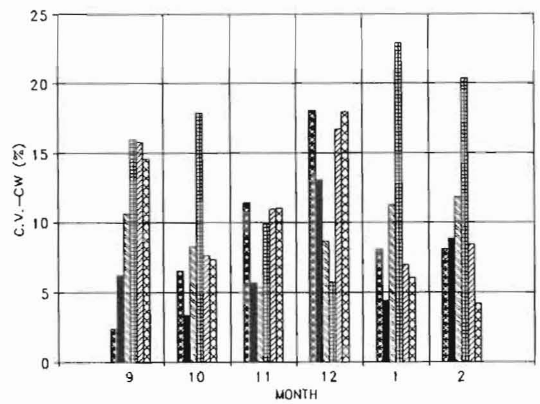

Figure 11. Histograms showing how values of the Coefficient of Variance vary from month to month when one month's chilled water data is used to identify a linear model from MRA and then using this model to predict energy use for the other months. Each of the six histograms for each month represents one month.

There is a large month-to-month variation in these values. For example, the September model which has a CV of 2.35%, predicts December values to a CV of about 16% only. Most models seem to predict the December values poorly, probably because this is the coldest month coupled with long abnormal building operation due to holiday schedules.

Figure 12 graphically depicts the tracking ability of the September model vis-a-vis future values.

Proceedings of the Eighth Symposium on Improving Building Systems in Hot and Humid Climates, Dallas, TX, May 13-14, 1992

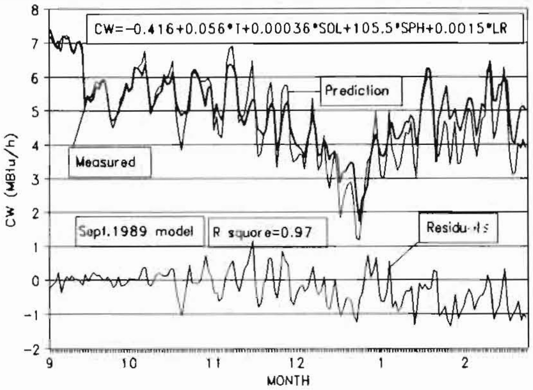

Figure 12. A time plot to illustrate how well a chilled water model identified using September daily data predicts use during the next five months. The corresponding residual time plot is also shown.

Even for December, the model does not seem to do too poorly. Part of the reason why the CV values for December are high (Figure 11) is that the mean values of CW are low for this month, which increases the CV values even when the RMSE are the same. However, if this month were not considered, CV values for CW are always less than 15% and mostly between 5-10%. This range of uncertainty values is quite satisfactory for most engineering applications.

CONCLUDING REMARKS

The primary objective of this paper was to evaluate MRA and PCA approaches as statistical means of identifying robust and accurate statistical linear models for predicting daily chilled water and hot water use in commercial buildings. Despite multicollinearity among the regressor variables and contrary to a recent study, we conclude, based on six months of measure data from a large engineering center located in central Texas, that a model identified by MRA has a slightly higher predictive power than one identified by PCA. Though there was no advantage in resorting to PCA over MRA, based on the data used, we argue that a more careful and elaborate investigation (than was considered necessary previously) using multiple data sets exhibiting a wide range of multicollinearity strengths is required in order to satisfactorily resolve this issue.

The paper also illustrates a procedure, based on generating synthetic building energy use data from a large simulation code, whereby one could ascertain whether influence of a particular variable on energy use was linear or not. This issue is fundamental to our approach of using linear regression for statistical model identification.

Finally, the predictive power of a regression model identified from one month's data has been evaluated by noting how well it fits for the remaining five months. The difference between model-predicted and measured values has been quantified in terms of monthly coefficient of variation values. These values have been found to be less than 10% for chilled water models.

ACKNOWLEDGMENTS

This research was funded and supported by the State of Texas Higher Education Coordinating Board as part of the ERAP Program and by the Governor's Energy Office, as part of Texas A&M's LoanSTAR Monitoring and Analysis contract. Special thanks are due to Dr. David Ruch for his helpful advice and insightful comments. We also acknowledge the assistance of Doug Bronson and Dr. Jeff Haberl for supplying us with the necessary DOE-2 simulation data.

REFERENCES

Bronson, D. J., Hinchey, S. B., Haberl, J. S. and O'Neal, D. L., 1992, "A Procedure for Calibrating the DOE-2 Simulation Program to Non-Weather Dependent Measured Loads", Proceedings of the ASHRAE Transactions, V. 98, Pt. 1, Anaheim, California.

Chen, L. L., "Developing a Change-Point Principal Component Predictive Model for Energy Use in a Supermarket," M. S. Thesis, Mechanical Engineering Department, Texas A&M University, College Station, TX.

Claridge, D., Haberl, J., Katipamula, S., O'Neal, D., Ruch, D., Chen, L., Heneghan, T., Hinchey, S., Kissock, K., and Wang, J., 1990. "Analysis of Texas LoanSTAR Data", Proceedings of the Seventh Annual Symposium on Improving Building System in Hot and Humid Climates, Texas A&M University, College Station, Texas, October 9-10, pp. 53-60.

Daultrey, S., 1976. Principal Components Analysis, Geo Abstracts, University of East Anglia, Norwich.

Draper, N., and Smith, H., 1981. Applied Regression Analysis, 2nd Edition, John Wiley & Sons, New York.

Fels, M. (Ed.), 1986. "Special Issue Devoted to Measuring Energy Savings, The Princeton Scorekeeping Method (PRISM)", Energy and Buildings, Vol. 9, Nos. 1 and 2.

Flury, B., 1988. Common Principal Components Related Multivariate Methods, John Wiley & Sons, New York.

Haberl, J. and Claridge, D., 1987. "An Expert System for Building Energy Consumption Analysis: Prototype Results", ASHRAE Transactions, V. 93.

Jackson, J. E., 1991. A User's Guide to Principal Components, A Wiley-Interscience Publication, John Wiley & Sons, New York.

Jollifer, I. T., 1986. Principal Component Analysis, Springer, New York.

Katipamula, S., and Claridge, D. "Use of Simplified Models to Measure Retrofit Savings," Proceedings of the ASME/JSES/KSES International Solar Energy Conference, Maui, Hawaii, March 1992.

Kissock, J. K., Claridge, D. E., Haberl, J. S. and Reddy, T. A., 1992. "Measuring Retrofit Savings for the Texas LoanSTAR Program: Preliminary Methodology and Results", Proceedings of the ASME/JSES/KSES International Solar Energy Conference, Hawaii, March, 1992.

LBL, 1981. DOE-2 Reference Manual, Lawrence Berkeley Laboratory and Los Alamos National Laboratory, LBL Report No. LBL-8706, Rev. 2, 1981; DOE-2 User Coordination Office, LBL, Berkeley, California.

Ruch, D., Chen, L., Haberl, J. and Claridge, D., 1991. "A Change-Point Principal Component Analysis (CP/PCA) Method for Predicting Energy Usage in Commercial Buildings: The PCA Model", Proceedings of the Seventh Symposium on Improving Building Systems in Hot and Humid Climates, Fort Worth, Texas, October, 1991.

Ruch, D. and Claridge, D., 1991. "A Four Parameter Change-Point Model for Predicting Energy Consumption in Commercial Buildings", Proceedings of 1991 ASME/JSES/JSME International Solar Energy Conference, Nevada, March, 1991.

SAS 1989, "SAS Language and Procedures: Version 6, 1st Ed., Cary, North Carolina, SAS Institute, Inc.

Sonderegger, R. C., 1978. "Diagnostic Tests for Determining the Thermal Response of a House", ASHRAE Transactions, Vol. 84, Part 1, pp. 691-702.

Subbarao, K., 1988. "PSTAR-Primary and Secondary Terms Analysis and Re-normalization: A Unified Approach to Building Energy Simulations and Short-Term Monitoring", SERI/TR-254-3175, Solar Research Institute, Golden, Colorado.

Van Rijckevorsel, J. L. A. and de Leeuw, J. D., 1988. Component and Correspondence Analysis. Dimension Reduction by Function Approximation, John Wiley & Sons, New York.

Verdict, M., Haberl, J., Claridge, D., O'Neal, D., Heffington, W., and Turner, D., 1990. "Monitoring $98 Million in Energy Efficiency Retrofits: The Texas LoanSTAR Program", Proceedings of the ACEEE 1990 Summer Study on Energy Efficiency in Buildings, ACEEE, Washington, D.C., August, pp. 7.261-7.271

Zaikong Chen, Jianxun Wu, and Jianno Shi, 1985. Computer Method for HVAC Engineering, Architecture Engineering Publishing House, Beijing, China.