Using synthetic data to evaluate multiple regression and principal component analyses for statistical modeling of daily building energy consumption

T.A. Reddy and D.E. Claridge
Energy Systems Laboratory, Texas A&M University, College Station, TX 77843 (USA)

(Received October 10, 1993; accepted in revised form February 28, 1994)

Abstract

Multiple regression modeling of monitored building energy use data is often faulted as a reliable means of predicting energy use on the grounds that multicollinearity between the regressor variables can lead both to improper interpretation of the relative importance of the various physical regressor parameters and to a model with unstable regressor coefficients. Principal component analysis (PCA) has the potential to overcome such drawbacks. While a few case studies have already attempted to apply this technique to building energy data, the objectives of this study were to make a broader evaluation of PCA and multiple regression analysis (MRA) and to establish guidelines under which one approach is preferable to the other. Four geographic locations in the US with different climatic conditions were selected and synthetic data sequences representative of daily energy use in large institutional buildings were generated in each location using a linear model with outdoor temperature, outdoor specific humidity and solar radiation as the three regression variables. MRA and PCA approaches were then applied to these data sets and their relative performances were compared. Conditions under which PCA seems to perform better than MRA were identified and preliminary recommendations on the use of either modeling approach formulated.

1. Background

Energy use in an increasing number of large commercial and institutional buildings is being continuously monitored in order to (a) measure retrofit energy savings, (b) measure retrofit demand savings and (c) avoid energy wastage by identifying and rectifying operation and maintenance problems and oversights as soon as they appear (for example, see refs. 1 and 2). A key component in such energy conservation programs is the ability to develop accurate models of energy consumption in a specific building. A case in point is the finding by Greely *et al.* [3] that of 1700 buildings in which energy retrofits were performed, about one in six had *predicted* savings within 20% of the measured values. For monitoring applications of types (b) and (c), hourly models are required. Several studies have addressed this problem; these studies could be broadly classified as:

(i) calibrated detailed computer simulations [4, 5];

(ii) calibrated simplified system modelling [6];

(iii) trigonometric or Fourier series analysis [7, 8];

(iv) statistical linear regression techniques [9, 10];

(v) macroscopic dynamic inverse modeling [11, 12];

(vi) artificial neural network [13].

However, for applications of type (a) above, it has been found that models on a daily time scale are generally (though not always) more appropriate [14]. Though shorter time periods may appear desirable, the reliability of data is enhanced by averaging/summing over longer time periods. Also, not having to deal with the strong diurnal fluctuations in usage patterns of commercial buildings makes the daily modeling approach simpler and therefore more attractive than hourly modeling.

The most widely used daily modeling approach is multiple linear regression analysis. This involves identifying a least-squares regression model between the daily energy consumption (which could be either the whole-building electricity use or the thermal energy use in the form of chilled water and hot

SSDI 0378-7788(94)00873-4

water or steam) and predictor variables (usually climatic variables, like mean daily outdoor temperature, humidity, solar insolation). Note that the effect of scheduling in commercial buildings is so pronounced that weekdays generally need to be distinguished from weekends and holidays and separate models identified for each type of period [14]. Multiple regression analysis (MRA) is relatively simple to understand, easy to implement and quite effective for many purposes [15]. The significant collinearity between the predictor variables themselves is, however, likely to lead to two different problems [16]:

(i) though the model may provide a good fit to the current data, its usefulness as a reliable predictor of future consumption is suspect. The regression coefficients tend to have large standard errors which makes the model unstable;

(ii) the regression coefficients in the model may no longer be proper indicators of the relative physical importance of the regressor parameters.

Building data analysts and engineers would like to circumvent such problems. Though several studies have acknowledged the dangers of multicollinearity and suggested analysis tools to avoid them, the relative importance of multicollinearity effects themselves and the effectiveness of the remedial action taken are difficult to evaluate properly because a valid basis of comparison is unavailable for dealing with measured building energy data. Draper and Smith [15] state 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. Mullet [17], discussing why regression coefficients in the physical sciences often have wrong signs, quotes (i) Marquardt who postulated that multicollinearity is likely to be a problem only when correlation coefficients among regressor variables are higher than 0.95, and (ii) Snee who used 0.9 as the cut-off point. In any case, researchers in building energy seem to assume that multicollinearity effects are likely to be problematic only when correlation among regressor variables is extremely high, though how much "high" is remains ambiguous.

2. Literature review

A statistical technique useful for describing and summarizing data which has the potential to overcome these difficulties is principal component analysis (PCA) (see, for example, ref. 18). PCA is a classic multi-variate technique developed by Pearson in 1901 and used by Hotteling in 1933 for analyzing covariance and correlation structuring [19]. It has since become increasingly popular in multi-variate statistical theory and is used to overcome multicollinearity effects. In essence, PCA takes a group of n variables and re-expresses them as another set of n variables, each of which represents a linear combination of the original variables while retaining all the information found in the original regressor variables. These indices, known as principal components (PCs) have several useful properties: (i) they are uncorrelated with one another, and (ii) 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. Consequently, the number of regressor variables in the model can be reduced with little loss in model goodness-of-fit. The coefficients of the PCs retained in the model are said to be more stable and, when the resulting model with the PC variables is transformed back in terms of the original regressor variables, the coefficients are said to offer more realistic insight into how the individual physical variables influence the response variable. Whether such features also apply to our problem of modeling building energy use is still uncertain. Nonetheless, the advantage of applying PCA to building energy modeling is not so much in its ability to summarize data, i.e., to reduce the number of regressor variables, but rather in its ability to remove the multicollinearity effects in the regressor variables (via the PCs) and thereby determine the importance of each of the climatic variables on energy use. It must be pointed out that, to date, PCA has so far been applied largely in such areas as the social sciences, where, lacking proper physical insight, models are generally weak and numerous correlated regressor variables tend to be included in the model [15].

Attempts to use PCA for building energy models have been few. Hadley and Tomlich [20] have investigated the relationship between daily residential space heating in four residences and six meteorological variables using PCA. They found that the first component, dominated by outdoor dry-bulb temperature, accounted for over 65% of the variation and that the first three principal components explained over 95% of the variation. Pearson and Palmiter [21] evaluated this approach as a tool for providing compact representation of monitored hourly energy use in an office building over an entire year. Hull and Reddy [22] used PCA as a means of identifying and clustering residential customers based on their air-conditioner-use profiles

during the hottest days of summer. Ruch *et al.* [16] also applied this approach to model whole-building daily electricity consumption in a grocery store in Texas and concluded that PCA appeared to produce a more reliable and physically plausible model than MRA. Cox [23] has further substantiated this claim by showing, based on engineering calculations of heat losses [24] in the same grocery store, that PCA captures the influence of individual physical variables on energy use more accurately than does MRA. Finally, a study by Wu *et al.* [25] of measured energy use in a large institutional building showed, however, that the PCA model did no better as a predictor model than an MRA model.

3. Scope

The above studies were of limited generality because data sets were relatively short (less than six months except for the study by Cox [23]), were limited to one geographic location, and used measured data along with all the inherent uncertainty unavoidably present. The objective of this study is to make a more generalized evaluation of both the MRA and PCA approaches using model-generated synthetic data representative of daily energy use in large institutional buildings. Specifically, we shall (i) compare the reliability with which two modeling approaches can determine the model coefficients, and (ii) evaluate their ability to accurately predict future energy use with the purpose of identifying conditions and determining guidelines under which one approach is superior to the other.

Using synthetic data in order to evaluate the soundness of particular parameter identification schemes is a widely used technique in several engineering disciplines. It has also been used in a few building energy studies (see for example, ref. 26). The advantage of using such pseudo-data is that the "correct" model coefficients of the regressor variables are known exactly, thereby providing a basis for meaningful evaluation. Another advantage of using synthetic data is that "noise" can be eliminated. In other words, the effects of numerous and unaccounted secondary physical influences can be removed from the model and we have a "clean" or idealized data set on which to evaluate our two modeling approaches. Using synthetic data can thus be likened to performing controlled experiments on a piece of equipment in a laboratory before installing it in the field.

4. Generation of synthetic data

The synthetic data sequences are generated as follows: (a) a hypothetical but realistic linear model of daily energy use in a large institutional building was assumed; (b) a year-long sequence of daily energy consumption using actual daily climatic data of a particular location was then generated; (c) random "noise" was finally introduced in the model to simulate real-world "noise". Three regressor variables most likely to influence energy use (E) in buildings are the outdoor dry-bulb temperature (TA), the specific humidity (SPH) and the solar radiation (SOL) (which in the present study shall be taken simply as the global horizontal radiation). The daily model used to generate the synthetic sequences was chosen as:

$$E = a_0 + a_1 \times \mathrm{TA} + a_2 \times \mathrm{SPH} + a_3 \times \mathrm{SOL} + k \times \epsilon \qquad (1)$$

where ϵ is a normally distributed random number with zero mean and unit standard deviation and k is a multiplier used to simulate varying levels of noise in the model. Thus, a poor model, characterized by a low R^2 value, can be generated by choosing a relatively high value of k, and vice versa.

The types and strengths of collinearity between the three climatic variables on a daily level are bound to influence the conclusions of the study. Hence in order to retain the general nature of this study the following four locations were chosen because of their widely different climatic characteristics:

- Albuquerque, NM — sunny and dry inland location;
- Fort Worth, TX — hot, sunny and humid inland location;
- Miami, FL — hot and humid coastal location;
- Seattle, WA — mild coastal location.

Hourly climatic data in the form of typical meteorological year (TMY) data [27] are available for the four sites. For each location, we have aggregated the hourly data into daily data and used this set with varying levels of the noise term to generate two year-long sequences of daily energy use. In order to simplify the analysis, the effects of weekdays, weekends and holidays have not been distinguished. The model coefficients and the specific values of k used to generate the sequences are given in Table 1. Note that the values have been determined by trial and error for each site such that a subsequent MRA to the annual data set would yield R^2 values of about 0.45 (representative of a poor model), and about 0.9 (good model). The model coefficients are hypothetical but considered realistic in the sense that the pattern and magnitude

TABLE 1. Values of k of eqn. (1) and model R^2 values used to generate the synthetic sequences. Model used:

$E = 10 + 2.0 \times \mathrm{TA} - 1.5 \times \mathrm{SPH} + 1.0 \times \mathrm{SOL} + k \times \epsilon$

where E is in GJ/day, TA in °C, SPH in g/kg and SOL in MJ/m^2 day

	Good model		Poor model	
	k	Model R^2	k	Model R^2
Albuquerque	35	0.900	120	0.453
Fort Worth	35	0.907	120	0.463
Miami	20	0.881	55	0.495
Seattle	22	0.916	80	0.460

of energy use E is similar to the daily chilled water energy use in a monitored institutional building in central Texas [25].

5. Preliminary data analysis

Figures 1 (a) and (b) are time plots of E (good model: synthetic sequence with high model R^2), and the individual contributions to the total energy use of TA, SPH and SOL for Albuquerque and Miami respectively, which represent the extremes in seasonal and day-to-day variation among the four locations selected. For example, the TA time plot is indicative of not the variation in temperature itself, but of $(a_1 \times \mathrm{TA})$ where a_1 is the model coefficient of eqn. (1). The plots give a visual indication of the relative importance of the three climatic variables on energy use specific to that location. Table 2 provides numerical values of the yearly mean contributions of the three driving forces for all four locations. We note from Table 2 that the model coefficients are such that temperature has the largest influence (about 40–50%). The solar influence is larger than the humidity influence in Albuquerque, Fort Worth and Seattle, the difference being pronounced in Albuquerque. These trends, considered physically realistic, are an indication that the model coefficients of eqn. (1), though arbitrary, are nevertheless realistic.

In terms of seasonal behavior, two separate trends are to be noted from Fig. 1. The energy use in Albuquerque (as well as Fort Worth which is not shown) shows the greatest variation over the year, while Miami has the least variation; a direct consequence of how the three climatic variables vary over the year. Also, while Albuquerque (as well as

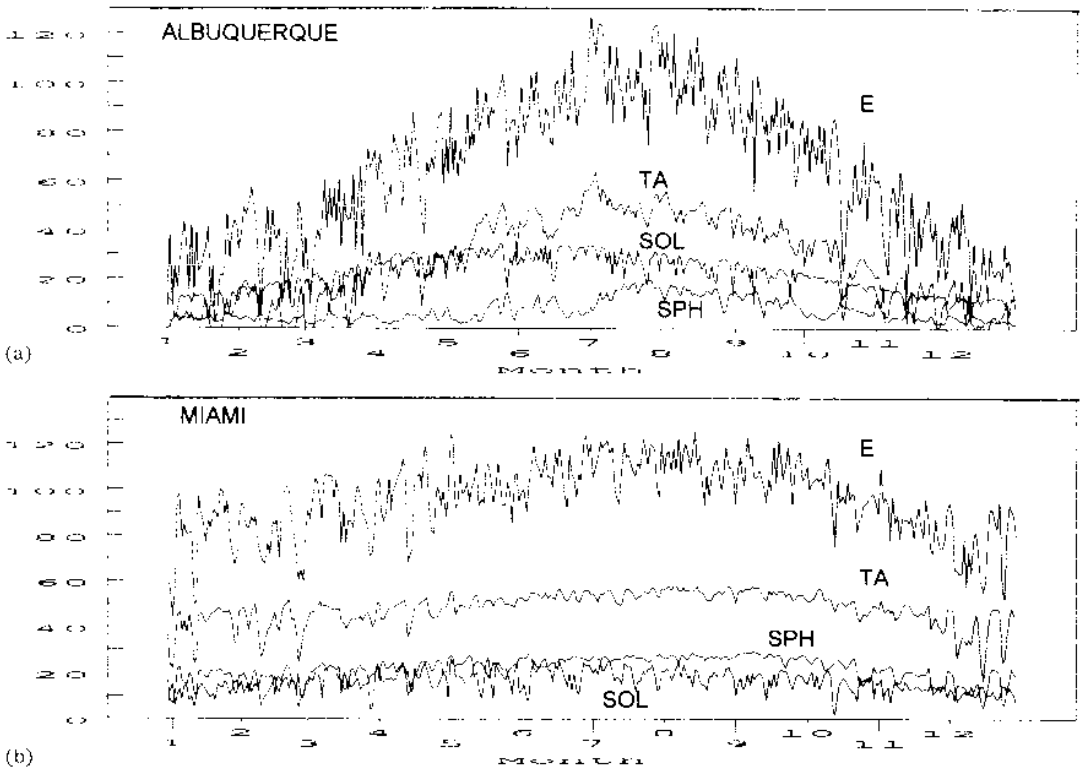

Fig. 1. Time plots of daily energy use (E) and the individual contributions of outdoor temperature (TA), specific humidity (SPH) and horizontal solar radiation (SOL) on E following eqn. (1) and Table 1 for two of the four locations chosen.

TABLE 2. Yearly mean relative contribution of the climatic drivers on energy use following eqn. (1)

	TA	SPH	SOL
Albuquerque	0.402	0.108	0.318
Fort Worth	0.470	0.183	0.217
Miami	0.501	0.221	0.175
Seattle	0.401	0.175	0.229

Seattle and Fort Worth which are not shown) seem to have day-to-day fluctuations of the same magnitude over the year, Miami exhibits important differences in seasonal fluctuations. The Dec.–Mar. period shows larger fluctuations while the day-to-day variations during the remaining periods are quite constant.

In order to explicitly study the effect of such variations over the year, we have divided the climatic data sets into four seasonal groups: Jan.–Mar. (winter months), Apr.–June (spring months) July–Sept. (summer months) and Oct.–Dec. (fall months). Though the seasonal breakup is arbitrary to some extent, it will not effect the conclusions of this study. Figures 2(a)–(d) show the mean and standard deviations of the three variables for each location. As is expected, TA and SPH values at Miami are generally higher than the other three locations and exhibit the smallest fluctuations (i.e., standard deviation values are lowest). Also, the variations in TA and SPH during the year are highest at Albuquerque while Seattle seems to have the largest annual relative variation as well as the highest day-to-day fluctuations in SOL. We note that generally the values of the three climatic variables during the seasonal periods of Apr.–June and Oct.–Dec. are closest to the annual values.

The correlation coefficients of the three climatic variables on both a seasonal and an annual basis have been computed and are shown in Fig. 3(a)–(d).

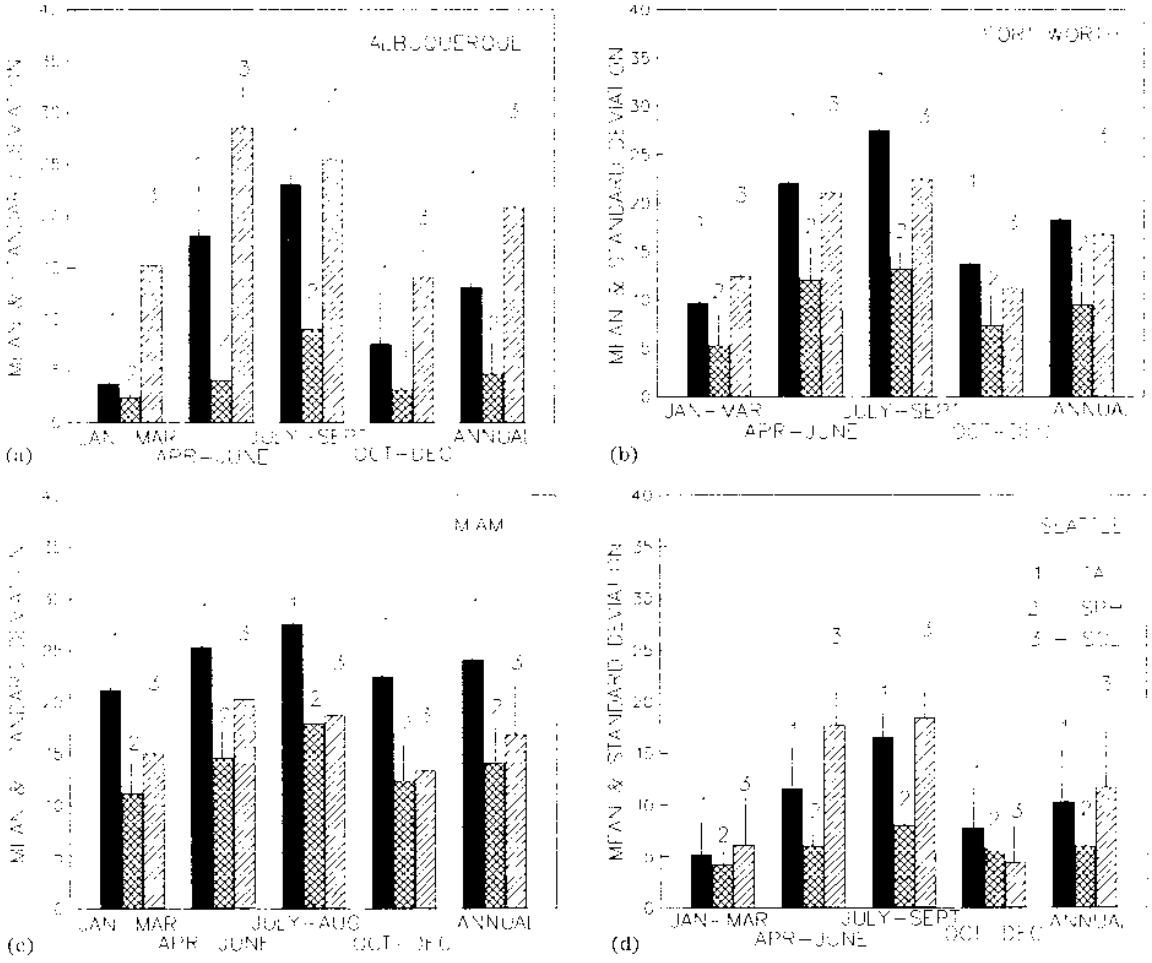

Fig. 2. Mean and standard deviation (shown as a whisker) of the three climatic variables for seasonal and annual periods. Units of TA, SPH and SOL are °C, g/kg and MJ/m² day, respectively.

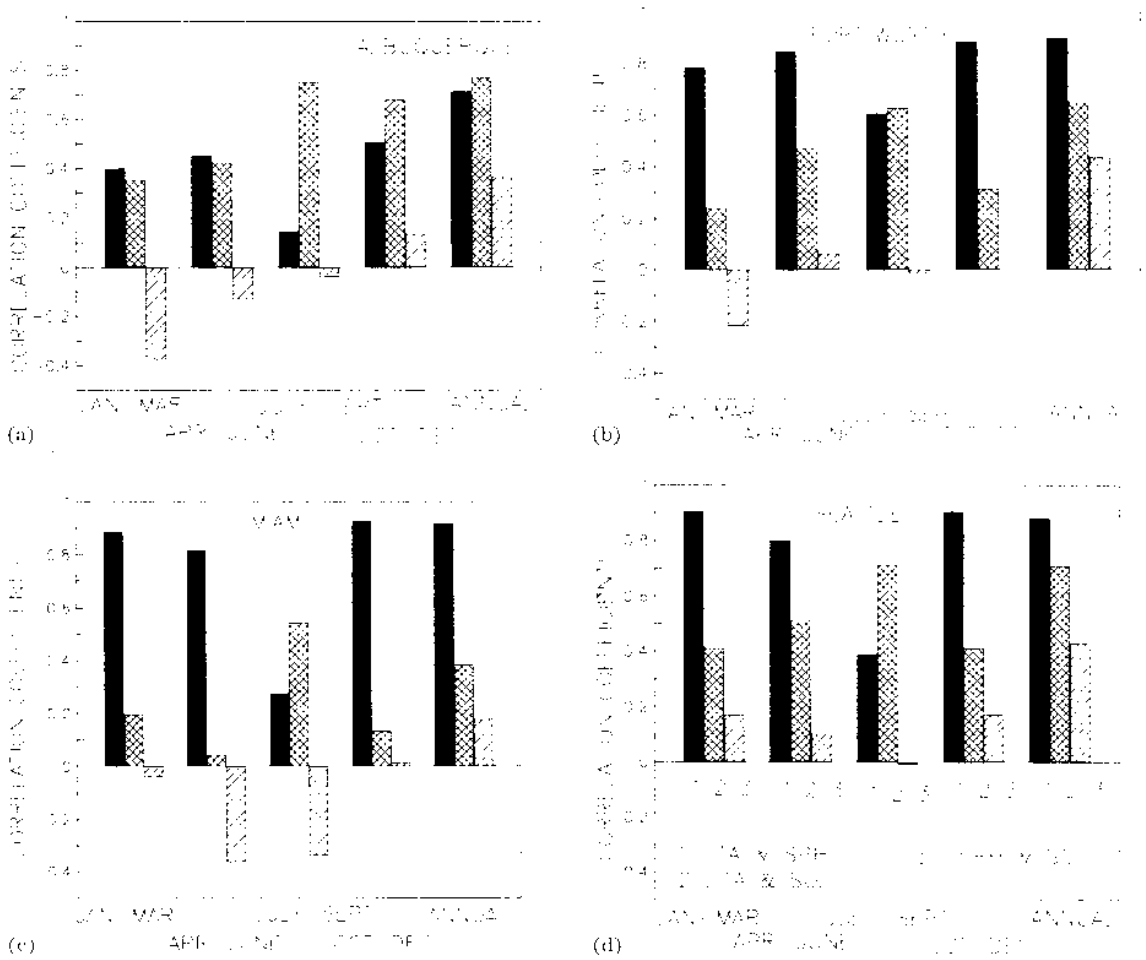

Fig. 3. Daily correlation coefficients for seasonal and annual periods.

For discussion purposes, correlation coefficients less than 0.4 can be considered weak while those greater than 0.7 can be considered strong. For all four locations, the summer months have distinctly different correlation patterns from the other seasons which in turn are generally similar. Except for Albuquerque, TA and SPH are strongly positively correlated (about 0.8 or higher) for all seasons other than summer. In summer, TA and SOL tend to be more collinear than TA and SPH. Climatic variables in Albuquerque are generally uncorrelated except for TA and SOL in the latter half of the year. The two variables also exhibit mediocre correlation strengths in Seattle and Fort Worth. The positive collinear behavior between TA and SPH on the one hand and TA and SOL on the other partly explains the fact that multiple regression of energy use versus climatic regressors often identifies "temperature" as being more influential than what one would expect based on physical considerations alone [14, 27]. A final noteworthy feature is that correlation strengths on an annual basis between the three climatic variables are generally as strong, if not stronger, than those during any one season.

6. The PCA approach

We shall briefly describe the various steps involved in using the PCA technique to our specific problem. The reader may refer to more standard textbooks such as Draper and Smith [15], Jollife [18], or Manly [19] for a more detailed and formalized mathematical description of the PCA approach.

(a) Since TA, SPH and SOL are physical variables which are measured on different and arbitrary scales, the data have to be "centered and scaled", i.e., standardized with a mean of zero and standard deviation of unity.

(b) The PCs of our standardized regressor data set are next determined using a commercial statistical package and ranked according to their var-

iance rank, i.e., their ability to explain the variance in the data set. A PC with a sufficiently low variance rank can be eliminated from the data set without losing a significant amount of information. Specifically, this means that the standard errors of the regressor coefficients of the PCs are lower than those of the MRA coefficients, albeit at the expense of a little lowering in the R^2 value of the model. Note that retaining all the PCs would result in a transformed regression equation identical to the MRA model. In our specific data sets involving all four locations, we have found that the first PC explains about 60–70% of the total variance, while the first two PCs explain over 90–95% of the total variance. This finding is consistent with that of Hadley and Tomich [20]. It should be pointed out, however, that the variance rank criterion of retaining the two influential PCs does not necessarily result in the best "predictive" model. We have found that the goodness-of-fit, i.e., model R^2 values between E and different sets of PCs, is a better criterion for selecting a particular set of PCs. This criteria has been adopted throughout this study.

(c) A multiple linear regression of energy use versus the reduced set of PC models is then performed; and finally,

(d) the regression coefficients of the PC models are transformed back in terms of the physical regressors using simple algebra. The resulting model of E in terms of TA, SPH and SOL is called the transformed PCA model.

It is clear that PCA is more demanding in time, effort and understanding than the MRA approach. However, it is systematic and can be conveniently programmed. Hence complexity is not so much the issue in the present comparison of PCA and MRA. Rather, we would like to identify conditions under which one approach is preferable to the other in terms of predictive accuracy and robustness of parameter estimates.

7. Inter-comparison

As noted earlier, the summer season for all four locations is noticeably different than the other seasons in terms of the correlation coefficient strengths of the three regressor variables. Since a proper comparison of MRA and PCA should explicitly consider such differences, we have identified MRA and back-transformed PCA models during the winter, summer and fall seasons separately, for both the low and high R^2 data sets. Subsequently, these models have been used with the entire year-long regressor data set to predict energy use during the entire year.

How well the MRA and PCA models predict yearly mean daily energy use can be gauged from Tables 3 and 4 for the high and low R^2 data sets respectively. In an effort to study model variance, annual root mean square errors (RMSE) of the residuals between model predicted and "measured" values of daily energy use were also computed for various model sets. These values are not presented here since no additional insight was evident. Figures 4 and 5 illustrate the values of the model coefficients of eqn. (1) as determined by the MRA and PCA approaches when applied to the winter, summer and fall subdata sets both for the high R^2 and the low R^2 data sets respectively. The values of the coefficients have been normalized by dividing them by their "correct" value given in Table 1. Hence a value of, say, 1.25 would imply that the statistical method overpredicts that particular coefficient by 25%. From the Figures, we note that the scatter of the normalized coefficients about unity is more pronounced for the low R^2 data set (i.e., Fig. 5) than for the high R^2 data set (i.e.,

TABLE 3. Yearly mean daily energy use predictions from MRA and PCA models identified from various subsets for the high R^2 data

Location	Approach	"Correct" value	Model identified from:		
			Jan.–Mar.	July–Sept.	Oct.–Dec.
Albuquerque	MRA	64.4	67.6	61.4	72.4
	PCA	64.4	68.7	65.7	72.5
Fort Worth	MRA	77.6	81.0	76.1	78.3
	PCA	77.6	80.4	76.5	75.3
Miami	MRA	96.1	95.6	95.6	95.8
	PCA	96.1	96.2	98.2	95.9
Seattle	MRA	51.1	51.4	50.4	50.2
	PCA	51.1	51.6	49.6	50.5

TABLE 4. Yearly mean daily energy use predictions from MRA and PCA models identified from various subsets for the low R^2 data

Location	Approach	"Correct" value	Model identified from:		
			Jan.–Mar.	July–Sept.	Oct.–Dec.
Albuquerque	MRA	64.5	75.7	63.2	67.3
	PCA	64.5	76.3	62.8	65.7
Fort Worth	MRA	77.6	89.1	72.6	80.0
	PCA	77.6	82.8	78.7	79.5
Miami	MRA	96.1	94.9	94.4	95.1
	PCA	96.1	95.8	106.1	95.9
Seattle	MRA	51.1	51.6	48.5	48.1
	PCA	51.1	51.2	44.8	47.7

Fig. 4). This merely supports the well-known fact that statistical determination of the model coefficients is likely to be poorer if the model itself is poor. However, even a superficial glance indicates that the bias in regressed coefficients from the "correct" values seems as important for the PCA models as for the MRA models.

Close inspection of the data in Tables 3 and 4 and Figs. 4 and 5 along with the correlation patterns shown in Fig. 2 lead us to the following salient observations:

(a) Differences in the predicted and "correct" yearly mean daily energy use are less than about 5% for the high R^2 data set (except for Albuquerque), while they are often larger for the low R^2 data set.

(b) If estimation of the model parameters of one approach is superior to the other approach, the prediction accuracy of that approach is usually better (it is so in three-fourths of the instances), and seems independent of the model R^2 of the data set. Hence, the two specific traits used to compare MRA and PCA in this study can be assumed to be generally consistent with each other.

(c) For Albuquerque, correlations between climatic variables are low during the season of Jan.–Mar., less than 0.4. Consequently, it is not surprising that MRA does a better job than PCA at both parameter identification and prediction for both sets of R^2 data. Hence, a basic criteria for considering using PCA in the first place is that at least one couple among the regressor variables should be intercorrelated to more than about 0.5.

(d) Parameter identification and prediction accuracy of PCA is likely to be superior to MRA (albeit slightly in most cases) whenever the regressor variable data set exhibits a strong positive correlation (about 0.7 or larger) *only* between *one* set of variables. This observation seems to hold true irrespective of how well the MRA model fits the data, i.e., for both the high and low R^2 data sets. This trend is deduced from the results of summer data at Albuquerque, and winter and fall data at Fort Worth and Miami. There are, however, two exceptions to this trend: (i) MRA seems to be better for the low R^2 data set at Albuquerque in summer; and (ii) when correlation coefficients are negative, as between SPH and SOL at Miami during summer, this trend is seen to be true only for the low R^2 data set.

(e) Whenever correlation coefficients of about 0.4 or higher exist between more than one set of variables, PCA is superior to MRA in the case of the *low* R^2 *model*, a tentative threshold range for the MRA model R^2 being about 0.5 or less. This behavior can be noted during fall in Albuquerque, summer in Miami and for all seasons in Seattle; with the only exception being summer in Fort Worth. For the *high* R^2 *model*, the results of our study are inconclusive. There seems to be an even split in the number of instances where one statistical technique was superior to the other.

8. Summary and conclusions

The primary objective of this study was to evaluate MRA and PCA approaches using model-generated synthetic data representative of daily energy use in large institutional buildings. A simple linear model was used involving three climatic parameters only. Actual year-long daily climatic data from four widely different climatic locations were used to drive the model. The criteria of comparison were (i) the reliability with which the approaches could "re-identify" the model coefficients, and (ii) their ability to accurately predict future energy use. Both these criteria were found to be consistent with each other, which simplified the comparison. A perhaps im-

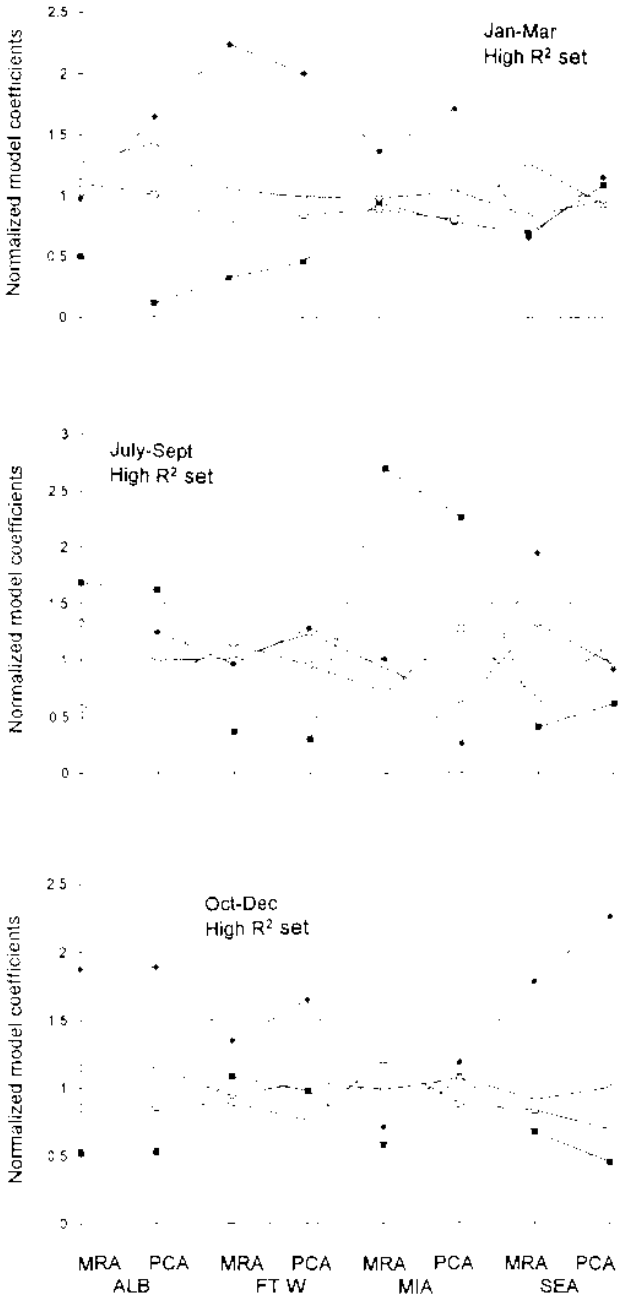

Fig. 4. Model coefficients of eqn. (1) identified by MRA and PCA for the high R^2 data sets and normalized by their corresponding "correct" values as given in Table 1. —■— a_0, -□- a_1, -◆- a_2, —◇— a_3.

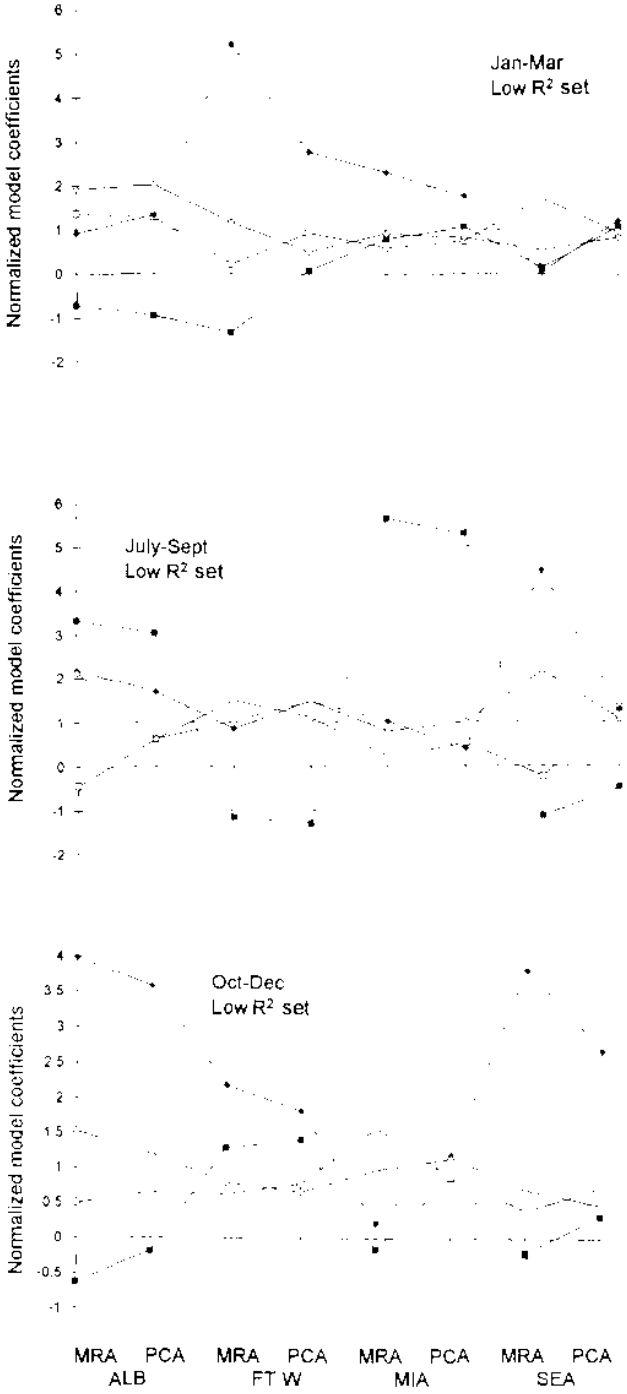

Fig. 5. Model coefficients of eqn. (1) identified by MRA and PCA for the low R^2 data sets and normalized by their corresponding "correct" values as given in Table 1. -■- a_0, -□- a_1, -◆- a_2, —◇— a_3.

portant issue that has not been investigated in this study is whether PCA regression coefficients have smaller standard errors than MRA coefficients. This evaluation would have required that the coefficients be re-identified under multiple sets of synthetic data with different random noise contributions, something not done in this study.

The results of this study indicate that use of PCA over MRA can be recommended only when, (i) at least one regressor variable set has correlation coefficient strengths of about 0.5 or higher, *and* (ii) the monitored energy data is poorly explained

by a MRA model, i.e., the model R^2 value is low (less than about 0.5); *or* when only one set of parameters shows significant collinearity (coefficients of 0.7 or more), irrespective of the model R^2 value. In any case, the improvement in using one approach as against the other is slight in most cases. The above recommendations should be viewed as preliminary indicators but are nevertheless deemed significant since injudicious use of PCA may exacerbate rather than overcome problems associated with multicollinearity.

Acknowledgements

Useful discussions and critical comments by D. Ruch and R. Cox are acknowledged. This study has been supported by the Texas Governor's Energy Office as part of the LoanSTAR Monitoring and Analysis Program.

References

1 J. Haberl and J. Vajda, Use of metered data analysis to improve building operation and maintenance: early results from two federal complexes, *Proc. ACEEE 1988 Summer Study, Asilomar, CA, 1988*, Vol. 3.

2 D. Claridge *et al.*, Improving energy conservation retrofits with measured savings, *ASHRAE J., 33* (10) (1991) 14–22.

3 K.M. Greely, J.P. Harris and A.M. Hatcher, Measured savings and cost-effectiveness of conservation retrofits in commercial buildings, *Rep. No. 27568*, Lawrence Berkeley Laboratory, Berkely, CA, 1990.

4 E. Hsieh, Calibrated computer models of communication buildings and their role in building design and operation, *Rep. No. 230*, Center for Energy and Environmental Studies, Princeton University, Princeton, NJ, 1988.

5 D. Bronson, S. Hinchey, J. Haberl, D. O'Neal and D. Claridge, A procedure for calibrating the DOE-2 simulation program to non-weather dependent measured loads, *ASHRAE Trans., 98* (Part 1) (1992) 1–5.

6 S. Katipamula and D. Claridge, Use of simplified systems model to measure retrofit energy savings, *ASME J. Solar Energy Eng., 116* (2) (1993) 57.

7 J.E. Seem and J.E. Braun, Adaptive methods for real time forecasting of building electrical demand, *ASHRAE Trans., 97* (2) (1991) 710–721.

8 A. Dhar, T.A. Reddy and D.E. Claridge, Improved Fourier series approach to modeling hourly energy use in commercial buildings, *Proc. ASME Int. Solar Energy Conf., San Francisco, Mar. 1994.*

9 J.R. Forrester and W.J. Wepfer, Formulation of a load predictor algorithm for a large commercial building, *ASHRAE Trans., 90* (Part 2B) (1984) 536–551.

10 M. MacDonald and D. Wasserman, Investigation of metered data analysis methods for commercial and related buildings, *Rep. No. ORNL/CON-279*, Oak Ridge National Laboratory, Oak Ridge, TN, 1988.

11 K. Subbarao, PSTAR - Primary and secondary terms analysis and renormalization: a unified approach to building and energy simulations and short-term monitoring, *Rep. SERI/TR-254-3175*, Solar Energy Research Institute, Golden, CO, 1988.

12 A. Rabl, Parameter estimation in buildings: methods for dynamic analysis of measured energy use, *ASME J. Solar Energy Eng., 110* (1988) 52.

13 J.F. Kreider and X.A. Wang, Artificial neural networks demonstration for automated generation of energy use predictors for commercial buildings, *ASHRAE Trans., 97* (2) (1991) 775–779.

14 J.K. Kissock, A methodology to measure retrofit savings in commercial buildings, *Ph.D. Thesis*, Mechanical Engineering Department, Texas A&M University, Dec. 1993.

15 N. Draper and H. Smith, *Applied Regression Analysis*, Wiley, New York, 2nd edn., 1981.

16 D. Ruch, L. Chen, J. Haberl and D. Claridge, A change-point principal component analysis (CP/PCA) method for predicting energy usage in commercial buildings: The PCA Model, *ASME J. Solar Energy Eng., 115* (2) (1993) 77.

17 G.M. Mullet, Why regression coefficients have the wrong sign, *J. Quality Technol., 8* (3) (1976).

18 I.T. Jolliffe, *Principal Component Analysis*, Springer, New York, 1986.

19 B.F.J. Manly, *Multivariate Statistical Methods: A Primer*, Chapman and Hall, London, 1986.

20 D.L. Hadley and S.D. Tomich, Multivariate statistical assessment of meteorological influences on residential space heating, *Proc. ACEEE 1986 Summer Study on Energy Efficiency in Buildings, Asilomar, CA, 1986*, Vol. 9.

21 E.W. Pearson and L. Palmiter, Issues in load shape representation, *Proc. ACEEE 1986 Summer Study on Energy Efficiency in Buildings, Pacific Grove, CA*, Vol. 9.

22 D.A. Hull and T.A. Reddy, A procedure to group residential air-conditioner load profiles during the hottest days in summer, *Energy, 15* (2) (1990) 105.

23 R.L. Cox, An analysis of grocery store energy use, *M. Sc., Thesis*, Mechanical Engineering Department, Texas A&M University, Dec. 1993.

24 ASHRAE, *Fundamentals*, American Society of Heating, Refrigeration and Air-conditioning Engineers Inc., Atlanta, GA, 1985.

25 J.X. Wu, T.A. Reddy and D.E. Claridge, Statistical modeling of daily energy consumption in commercial buildings using multiple regression and principal component analysis, *Proc. 8th Symp. Improving Building Systems in Hot and Humid Climates, Dallas, May 1992.*

26 A.K. Meier, J. Busch and C.C. Conner, Testing the accuracy of a measurement-based building energy model with synthetic data, *Energy Build., 12* (1988) 77–82.

27 M. Fels (ed.), Special issue devoted to measuring energy savings, The Princeton Scorekeeping Method (PRISM), *Energy Build., 9* (1/2) (1986).

28 I.J. Hall, R.R. Prairie, H.E. Anderson and E.C. Boes, Generation of typical meteorological years: 426 solmet stations, *ASHRAE Trans., 85* (1979) 507.