PROCEDURES FOR FILLING SHORT GAPS IN ENERGY USE AND WEATHER DATA

Hui Chen
Research Associate
Energy Systems Laboratory
Texas A&M University System
College Station, Texas
Phone: (979) 458-2656
Fax: (979) 862-2457
Email: chen@esl.tamu.edu

David E. Claridge, Ph.D., P.E.
Professor, Associate Director
Energy Systems Laboratory
Texas A&M University System
College Station, Texas
Phone: (979) 845-1280
Fax: (979) 862-2457
Email: claridge@esl.tamu.edu

ABSTRACT

Filling short gaps (a few hours) in hourly energy use and weather data can be useful for (i) retrofit savings analysis and calculation, and for (ii) diagnostic purposes. The paper evaluates four methods for rehabilitating short periods of missing data. Single variable regression, polynomial models, Lagrange interpolation, and linear interpolation models are developed, demonstrated, and used to fill 1-6 hour gaps in weather data, heating data and cooling data for commercial buildings. The methodology for comparing the performance of the four different methods for filling data gaps uses 11 one-year data sets to develop different models and fill over 50,000 "pseudo-gaps" which are created by assuming data is missing and then comparing the "filled" values with the measured values. Comparisons are made using six statistical parameters including mean bias error, coefficient of determination, and coefficient of variation of the root-mean-square-error.

For filling 1-6 missing hours of cooling data, heating data or weather data, a linear interpolation model or a polynomial model with hour-of-day (HOD) as the independent variable both provide a mean bias error of less than 0.087 % (0.005 F). The Lagrange model exhibits mean bias errors of 0.175 % (-0.010 F) which is better than the SVR model with temperature as the independent variable, which exhibits mean bias errors up to 0.909 % (0.062 F).

Based on these findings, the polynomial model with hour-of-day as the independent variable and the linear interpolation model are recommended for filling data gaps of six hours or less in cooling, heating and weather data.

INTRODUCTION

A successful building retrofit monitoring and analysis program depends on the collection of monitored building energy use data (at the daily and hourly level), screening data, determining retrofit savings and should include optimization of building energy use by improving operation and maintenance practices. There are large quantities of missing data or bad data in both the LoanSTAR database and the NWS weather database due to data processing problems or instrumentation and monitoring hardware problems. According to the statistical data from 1,020 channels of building energy data examined in 1993, roughly 2% of the data in the database cannot be restored and are noted as missing data. 6% of the data points have required some sort of correction after they were collected. Most of corrections involve the flow meter corrections (Haberl et al. 1993). The 1-6 hour data gaps cover all missing NWS temperature and dew point data, and 50-70% of total missing LoanSTAR temperature and humidity data, and 50-70% of total missing LoanSTAR energy use such as cooling, heating, motor control use and electricity data. Metered data analysis is a crucial aspect of any energy conservation program. However, it is hard to determine retrofit savings when there is insufficient monitored data (energy use or weather data) or there are large amounts of bad data. Failure to measure retrofit savings will hinder the adoption of efficiency measures in buildings. Filling in missing data in commercial buildings can be useful for: (i) retrofit savings analysis and calculations, (ii) diagnostic purposes, and (iii) acquiring physical insight into the operating pattern of buildings. Lots of missing (or bad) energy use and weather data is common in determining energy savings of retrofitting building and causes problem with the result of energy analysis (Claridge, et al., 1990a; Claridge, et al., 1990b).

The objectives of this research project was to: 1) check energy use and weather data quality and examine the missing data distributions, 2) demonstrate the effectiveness of selected methods for filling in missing data, 3) develop various modeling procedures using different approaches, 4) compare the results (the measured and predicted data in the same data sets) by means of statistical parameters, and 5) identify the criteria which is used to evaluate the methods. The purpose of this thesis is to identify simple and convenient method to rehabilitate missing data (energy use and weather data) for commercial building energy use evaluation.

A number of studies have been made on energy use data and weather data to determine statistical parameters of interest for thermal environmental engineering applications. Attempts at Fourier series modeling of hourly energy use in commercial buildings are relatively few, the most important being perhaps that by Steven and Raymond (1996) who chose a week as the maximum period of the Fourier series. The regression fit was poor, however, partly because of the choice of the maximum period. Dhar et al. (1995) suggested that the primary day-types be determined by using the calendar method and presented a generalized Fourier series approach which, while ensuring a wider range of applicability, also yields superior regression fits partly because of the care taken to separate the days of the year during which commercial buildings are operated differently, and partly because of the rational functional form of the regression model. Climatic data (i.e., solar radiation and outdoor temperature) are periodic and have been analyzed and simulated using the Fourier series by several researchers (Hittle and Pederson, 1981; Hokol, et al., 1990; McCutchan, 1979; Philios, 1984;). The shape of diurnal curves in the ambient air has been modeled in a variety of ways. These methods include simple curve fitting models based on sine curves (Allen, 1976; de Wit, 1978; Johnson and Fitzpatrick, 1977a, b) or Fourier analysis (Hokoi, et al., 1990; McCutchan, 1979) and more complex energy budget models (Floyd and Braddock, 1984; Akbari, et al, 1988). Since observed diurnal temperature curves are a combination of periodic sine and exponential decay curves, they are not readily represented by a few terms of a Fourier series. The sine-exponential model is used for a daily cycle of air temperature (Padit and Wu, 1983; Wann et al., 1985). The sine-exponential model uses a truncated sine function for the daytime and an exponential function at night. The sinusoidal model uses a cosine function to describe variation for the period from the time of minimum temperature to the time of maximum temperature and another cosine function for the period from the time of maximum temperature to the time of the minimum temperature in the next day.

A linear model assumes linear changes in temperature with respect to time from one extremum to the next. The mean time (with respect to sunrise) of minimum temperature and the mean time of maximum temperature are required parameters. This linear model was used to generate four years of hourly temperature using as input the daily highs and lows recorded by the National Weather Service at the Minneapolis-St. Paul International Airport for the years 1970, 1971, 1973 and 1974. The output from the model was compared with the actual hourly values recorded at the same site. Residual sums and standard errors computed for the model, both as functions of time of day and month of the year, showed that the model was more accurate in summer than in winter. Another linear model is used to calculate missing daily temperature data within stations in northern and central Idaho (Kemp, et al., 1983). The model uses the average of within station values recorded on the same day to replace a missing value, namely uses the data from the first day on either of the missing days. The moving average model (Kemp, et al., 1983) also belongs to the linear model. The model also uses the average of within station values recorded on the same day to replace missing data, which is same as the above linear model, but the model uses the data from two days on either side.

The Additive procedure (MDIF) model between stations uses the average of the estimates of the between station predictions to form a replacement value (Kemp et al., 1983). Kemp (1983) used a linear model as well as the Lagrangian polynomial model and a local spline to predict the temperature and geopotential height fields. The regression model used by Kemp et al. (1983) used weighted regression to determine replacement values where the replacement equations are the weighted sums of the between station regression equations. The model generated smaller errors when compared to the above linear methods.

Depending on the number of sequential records of missing wet bulb temperature data, DOE-2 either linearly interpolates between two available web bulb temperatures or fills in the last wet bulb temperature value in all subsequent records until the next record with data called the last data value. If the data gap is less than 24 data points, DOE-2 linearly interpolates, but if the missing data number is 24 points or greater, DOE-2 fills in the records with the last data. The DOE-2 weather packer also used the saturated

ambient condition method (Bronson, 1992). The DOE-2 weather packer compensates for the missing records by linear interpolation so that no abrupt changes result on the weather tape. The linearly interpolation and psychometric relationships are mainly used to fill in data gaps in TMY2 and DOE-2. Of course, these two methods are related to this paper and can be used as the interesting methods. Problem is hard to use psychometric relationships to rehabilitate energy data gap. General speaking, energy use data can not be simply represented by means of the relationships between some psychometric variables, but weather data are easy to make good use of the psychometric relationships.

In the literature of above meteorological and oceanic methods, most of these methods use a weather model (such as Fourier series model) to predict a whole weather temperature profile or solar profile based on monthly, seasonal or yearly averages. Other methods are used to predict some weather data by using interpolation between two weather stations. These methods fail to be precise enough for use where there are significant short-term interactions between the weather data. Some investigators have used the finite Fourier series approximation to model energy use or weather data. Some Fourier series models are based on the calendar method (data-type such as weekdays) and most of models involve multiple variables such as dry-bulb temperature, specific humidity, etc. This paper emphasizes the filling of short gaps in hourly data and the "neighborhood relation" between hourly adjacent available data and missing data. Hourly energy use and weather data do not have the smooth periodic patterns due to its periodic oscillation. Therefore, the Fourier series approach is not considered to be one of simple and convenient methods studied in this thesis.

The development of regression techniques and numerical modeling (linear and Lagrange interpolation) approaches with a single variable were identified as the topics of current interest. These approaches are paid great attention because the objectives of the research are to adapt a simple and convenient method to rehabilitate missing data, and these two methodologies have the potential to offer high prediction accuracy. They also take less time and require less detailed information than a calibrated deterministic approach, the Fourier series approaches and the multi-variable regression models. The multi-variable models have slightly higher accuracy than the single variable models; however, single variable models are much more popular than multi-variable models because it is often very difficult to obtain solar and dew point temperature data (Kissock, 1993). Furthermore, the neighborhood relation between adjacent available data and missing data is very useful information which contains missing data' value and pattern. Various models with single variable (either temperature or HOD) can make good use of the neighborhood relation to deal with data gaps. Therefore, only single variable models (single variable regression, polynomial, linear and Lagrange interpolation) are investigated in this study.

This paper presents a review of each model considered and evaluates four simple and convenient methods, which can be used to replace missing energy use and weather data for the LoanSTAR database. The methods examined are single variable regression, polynomial interpolation, linear interpolation, and Lagrange interpolation. These methods are examined with temperature and with hour-of-day as the independent variable and the accuracy of these models is compared. The measured data, which are adjacent to each side of a short data gap, are used to develop each model based on the assumption that this measured data contains adequate information on the data pattern of the missing data. The different methods for filling data gaps will be compared using the same statistical parameters (MBE, RMSE, CV-SAE and SSE) to evaluate the prediction error of each model (polynomial, Lagrange, linear interpolation and single variable regression).

METHODOLOGY

1. Method for Developing Various Models

The datasets examined are all time series records of weather or energy use data. Measured data are available for the time series records adjacent to the gaps of 1-6 hours investigated. Observation shows that these adjacent measured data points are directly related to or contain important information about the missing data pattern. The methods investigated will utilize the adjacent measured data points, namely use the neighborhood relation between adjacent measured data and data gap, to develop different models and then predict some missing data.

The methods to be described are evaluated for data gaps of 1-6 consecutive hours. The gaps evaluated are created by assuming a gap starting with the 13th hour of a one-year data set, filling in the data performed the evaluation; it is then assumed the gap begins with the 14th hour of the data set, the gap is filled, etc. Thus hours 13 to 13+n-1 are assumed missing for a gap of length n. The data gap is then filled using the method being evaluated.

These gaps will be called pseudo-missing hourly data gaps to denote their method of creation. Each of the models except the linear interpolation model and the Lagrange model use the 12 data points on each side of the pseudo-missing data (24 total points) to create a model and fill the gap. The linear interpolation model is based on a single measured point on either side of the data gap, and the Lagrange model is based on four measured data values on either side of the data gaps. The different models developed are based on the same 24 hourly-measured data set (12 hourly measured data points on either side of data gap)

2. Criteria Used to Evaluate the Methods

The criteria used to evaluate methods for filling data gaps will consist of two parts: model accuracy and model simplicity. Procedures, which are easy to understand and implement, are preferable.

The various models are evaluated and compared by statistical parameters. These parameters are the mean bias error (MBE), the coefficient of variation of the root mean square error (CV-RMSE), the coefficient of variation of the ratio of the sum of the absolute errors to the sum of measured data points (CV-SAE) and the coefficient of multiple determination (R^2). These statistical parameters are defined by the following equations, respectively:

The Coefficient of Variation of the Root Mean Square Error (CV-RMSE) is a nondimensional measure that is found by dividing RMSE by the mean value of y.

$$CV-RMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p}}}{\bar{y}} \times 100 \quad (1)$$

The CV-SAE is a non-dimensional measure of the average deviation (in percentage) from the model.

$$CV-SAE = \frac{\sum_{1}^{n}|y_i-\hat{y}_i|}{\sum_{1}^{n} y_i} \times 100 \quad (2)$$

The normalized Mean Bias Error for energy use is the following equation:

$$MBE = \frac{\sum_{1}^{n}(y_i-\hat{y}_i)}{\sum_{1}^{n} y_i} \times 100 \quad (3)$$

and for weather data MBE is defined by equation:

$$MBE = \frac{\sum_{1}^{n}(y_i-\hat{y}_i)}{(n-p)} \quad (4)$$

This parameter Coefficient of Determination (R^2) is defined as follows:

$$R2 = \left[1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(\bar{y}-y_i)^2}\right] \times 100 \quad (5)$$

where

$\hat{y}_i$ is the i^{th} predicted dependent variable value for the same set of independent variables,

y_i is i^{th} data value of the dependent variable (actual measured data),

$\bar{y}$ is the mean value of the dependent variable in the data set and

n is the number of data points in the dataset.

DIFFERENT MODELS

1. Polynomial Model

A one variable polynomial model is defined as follows:

$$y = a_0 + a_1 x + a_2 x2 + \cdots + a_m x^m + e \quad (6)$$

where y represents the dependent variable and x the independent variable. The highest exponent, or power, of x used in the model is known as the degree of the model, and it is customary for a model of degree m to include all terms with the lower power of the independent variable. The least square method is used to estimate of the parameters $a_0, a_1, a_2, \ldots, a_m$ that minimize the sum of the squared differences between the actual approximating y values and the values y predicted by equation (Steven, and Raymond, 1996; Erwin, 1983).

The optimum polynomial regression model based on actual measured data points is not just the polynomial model for which mathematical convergence of the model is achieved; in addition the optimum regression should not oscillate wildly. Since the actual measured weather and energy use data points have a physical basis in the diurnal cycle or diurnal operating schedule due to solar variation, the optimum polynomial model should be of the most suitable order or power and should be based on a suitable number of data points. The polynomial theory shows that the most suitable power of the polynomial model is determined by a smallest value of SSEs, but in real polynomial applications on actual data gaps, the most suitable power of the polynomial model is 8 for energy data and 10 for dry bulb temperature data and dew point. The determination of the best polynomial order was made by the following procedures:

- The optimum polynomial order was determined for all pseudo-gaps (the nine chilled water and hot water data sets analyzed in the paper; this progress was repeated for data gaps of 1-6 hours.
- The optimal order was averaged over all gaps of a given length for CW and HW, and averages of the model fitting errors (SAE, MAE, R^2, CV(RMSE), MBE, StdErr, CV-SAE and SSE) were calculated.
- Similar calculations were performed for eight years of weather.

In order to reduce the residuals or errors, the most suitable number of measured data points for use in model development needs to be determined. The performance of models with five points, six points, eight points, and 10 points, respectively, on either side of the data gaps, is compared with the performance of models with 12 points on either side of the data gaps. Each set of comparisons are performed by creating approximately 100,000 data gaps using the 11 years of energy data and creating and filling pseudo-gaps corresponding to all of the data points in each data set which have enough points on both sides of the pseudo-gap to create the model. The filled data is then compared with the actual data and averaged over each month and the monthly averages are then averaged to perform the comparisons. The eight comparisons shown for each model type are the mean bias error (MBE) in percent, root-mean-square-error (RMSE) in MMBtu/hr, sum of the squared errors (SSE) in $MMBtu^2/hr^2$, the coefficient of variation of the RMSE (CV) in percent, the sum of the absolute error (SAE), and the percent of the filled data points which are within 5%, 10%, and 15% of the actual values, respectively.

Careful examination of the data shows that the models with 12 points on either side generally show much better statistical performance than the models with 5, 8 or 10 points on either side, except for mean bias error. The models with 12 data points on either side are chosen due to the known diurnal cycle in ambient temperature and building schedules. The most suitable number of measured data points for the polynomial model development is around 24 measured points for cooling, heating data and NWS temperature data.

2. Numerical Approach

Numerical analysis provides a convenient method for obtaining an approximation for filling missing data. The accuracy of error in the computed result depends upon approximate data or approximate methods or both. These data of function $f(x)$ may be spaced either evenly or unevenly along time x. This requires a method for finding the missing data of $f(x)$ between the tabulated points (interpolation) or outside the range in time x of the tabulated points (extrapolation). Interpolation is the process of estimating a value of the function for any intermediate value of the variable by a procedure other than the law which is given by the function itself but rather than with the help of certain given values of the function correcting to a number of variable values, while extrapolation is the estimation for some such values which lie outside the given values.

For various numerical or experimental reasons it is often convenient or possible to use Lagrange interpolation at both equal and unequal intervals (Steven, C.C. and Raymond, P.C., 1996; Erwin K., 1983). The Lagrange interpolating polynomial can be represented concisely as

$$p_n(x) = \sum_{j=0}^{n} f(x_i) P_j(x) \qquad (7)$$

where $$P_j(x_k) = \{ \prod_{j \neq 0}^{n} \frac{x - x_j}{x - x_i}, k = j, 0, k \neq j \} \qquad (8)$$

If a function f(x) is known only on the interval $a \leq x \leq b$, but values of f(x) are needed for x<a or x>b, then extrapolation is required. Even under the best of circumstances, extrapolation contains a strong element of uncertainty.

The linear interpolation model proposed for data gaps is of the following form:

$$f1(x) = f(xo) + \frac{f(x1) - f(xo)}{x1 - xo}(x - xo) \qquad (9)$$

The linear interpolation used assumes linear changes in energy use and weather data with respect to time from one extreme to the next. Unlike interpolation, where the function is firmly anchored on both sides of the point where a value is to be obtained, in extrapolation the function is fixed on only one side and is relatively free to wander on the other side.

3. Single Variable Regression Models

The outside air dry-bulb temperature is taken to be the only regression variable. The model for weather independent use is one-parameter (1-P), and those for weather dependent use include two parameter (2-P) (Kissock, 1993), three parameter (3-P) (Fels, 1986), four parameter (4-P) (Ruch and Claridge, 1992; Kissock et al., 1992) and five parameter (5-P) (Reddy et al., 1997) models. The modeling techniques can be adapted to model hourly data. The functional forms of these models are as follows:

1-P model: $E = B_0$ (10)

2-P model: $E = B_0 + B_1(T)$ (11)

3-P model: $E_h = B_0 + B_1(B_2 - T)^+$ for heating

$E_c = B_0 + B_1(T - B_2)^+$ for cooling (12)

4-P model: $E_h = B_0 + B_1(B_3 - T)^+ - B_2(T - B_3)^+$ for heating (13)

4-P model: $E_c = B_0 - B_1(B_3 - T)^+ + B_2(T - B_3)^+$ for cooling (14)

5-P model: $E = B_0 + B_1(B_3 - T)^+ + B_2(T - B_4)^+$ for heating and cooling (15)

In these equations, the symbol $(\ )^+$ indicates that these quantities are set to zero when negative. B_0 is the mean energy consumption, or the energy consumption at the change point temperatures, B_2, B_3, B_4, B_1 and B_2 are the temperature slopes.

The applicable parameters as slope and/or consumption at change points for cooling and heating energy consumption. These linear and change point linear models have physical significance to the actual heat loss/gain mechanisms. A one-P mean model for each hour of the day may be used to model weather independent energy use, for example, lighting and equipment loads in commercial buildings. The 3-P models are appropriate for modeling envelope-driven energy consumption in buildings without simultaneous heating and cooling, such as multi-family housing and small commercial buildings. The 4-P models are appropriate for modeling heating and cooling energy use in variable-air-volume (VAV) systems and/or in buildings with high latent loads. These models are also appropriate for describing heating and cooling energy use caused by some hot deck reset schedules and economizer cycles, but other models are sometimes more suitable for these cases. The 5-P models are appropriate for modeling energy consumption data which includes both heating and cooling, such as whole-building electricity data from buildings with electric heat pumps or both electric chillers and resistance heating. These models are also appropriate for modeling fan electricity consumption in variable-air-volume systems. The basic difference between a 4-P and a 5-P model is that the former has only one change-point temperature while a 5-P model has two. Physically, one would expect a residence not to use either heating or cooling over a certain outdoor temperature range, and this behavior is well captured by the 5-P model (Reddy, et al., 1997).

The research on the effect of independent variables to building energy use indicates that the multiple regression provide some significant insights on building energy use (Haberl and Claridge, 1987; Katipamula et al, 1994; Katipamula et al, 1995; Kissock, 1993; Leslie et al 1986). Multiple regression can be used both for individual buildings and to study characteristics that lead to differences in energy consumption between buildings. These methods are important for advancing analysis approaches for commercial buildings and may be used to predict missing energy use. However, the multiple regression models need more both weather related and other factors (such as internal load, etc.) which normally are hard to obtain, so the multiple regression models are not applied to energy data gap rehabilitation as the simple and convenient method in this paper.

COMPARISON OF DIFFERENT MODELS

This part puts all selected models together and compares their predicted accuracy using the same

measured data sets. It then presents the most accurate model that is closest to the actual measured data.

1. Comparison of Polynomial Energy Use Models with Hour-of-Day (HOD) or Temperature as Independent Variables

In order to ascertain whether temperature or HOD is the prefered independent variable in a polynomial model to be used for filling in missing data, this investigation only applies to gaps in energy use data, since gaps in temperature data cannot be filled using temperature as the independent variable, and gaps in humidity data are often coincident with gaps in temperature data. Comparison will be made between the performance of polynomial models using temperature or HOD as the independent variable first with two short datasets selected to illustrate the ability of the two approaches to fill data gaps and then comparison is made with results from filling over 50,000 pseudo-gaps created in 11 one-year data sets.

Two kinds of polynomial models, which are based on the same short datasets, are developed and used to fill in 6 hourly consecutive pseudo-missing peak values and bottom values of cooling data. The short datasets used are the 48 hours of cooling data for July 21, 1992 – July 22, 1992 from the Zachry Building. For each 6-hour pseudo-missing period (peak and bottom), both kinds of polynomial models are used to produce the various error parameter values for the cooling data gap. Figure 1-2 show that the statistical parameters such as CV-RMSE and CV-SAE of the polynomial model with HOD as the independent variable are always lower than the ones of the polynomial model with temperature as the independent variable whether the data gap is a set of peak values or not. The total gap-filling errors of the polynomial models with temperature as the independent variable are likewise higher than the errors of the polynomial models with HOD as the independent variable.

Figure 3 shows the different graphs for filling in 6 consecutive hourly peak and bottom cooling data gaps for the Zachry Building. The polynomial model with HOD as the independent variable fits peak and bottom data gaps shown for cooling datasets better than the polynomial model with temperature as the independent variable.

While the short datasets show that HOD is clearly the preferred independent variable for the specific gaps evaluated, and strongly suggest that HOD is the preferred variable for general use, the comparison is now extended to use over 50,000 pseudo-gaps created in 11 one-year datasets.

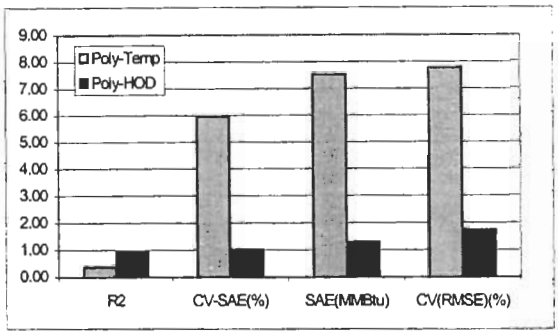

Figure 1 Comparison of Polynomial Model Parameters Based on One 6 Peak Gap

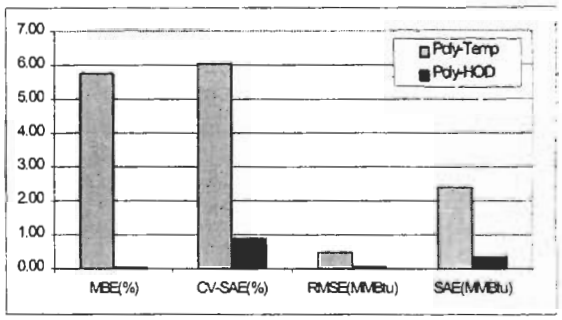

Figure 2 Comparison of Error Parameters Based on Different Polynomial Models

The following long cooling datasets are used to evaluate the cooling polynomial models:

- Two one-year chilled water datasets for the Zachry Building at Texas A&M during the periods from 9/14/1989-9/14/1990 and 12/20/1991-12/20/1992.
- Two one-year chilled water datasets for Taylor Hall at UT Austin during the periods from 6/22/1996-6/22/1997 and 7/17/1997-7/17/1998.
- One-year of chilled water data for the Geology Building at UT Austin during the periods from 2/1/1996-2/1/1997.
- One-year of chilled water data for the Main Building at UT Austin during the periods from 4/6/1993-4/6/1994.

The following long heating datasets are used to evaluate the heating polynomial models:

- Two one-year hot water datasets for the Zachry Building at Texas A&M during the periods from 9/14/1989-9/14/1990 and 12/20/1991-12/20/1992.

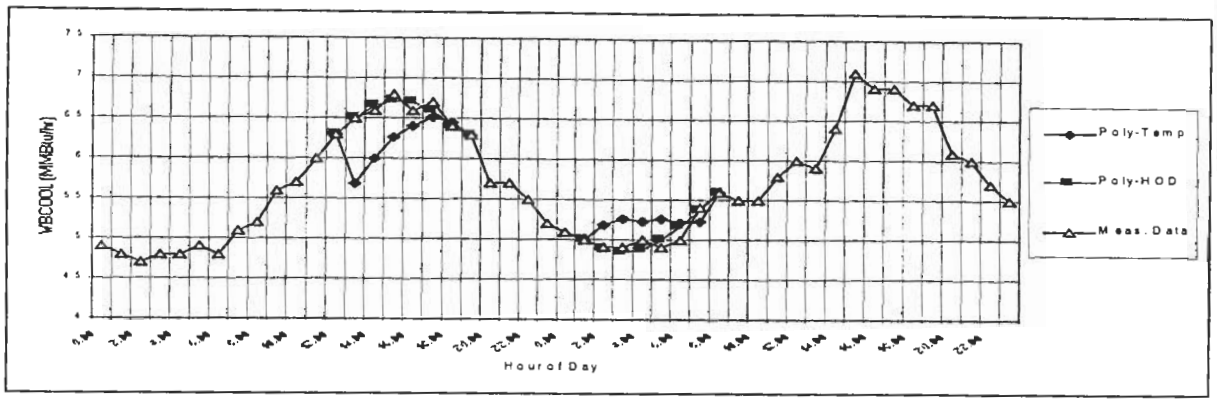

Figure 3 Comparison of Different Polynomial Models for Filling Peak and Bottom Data Gap

- Two one-year hot water datasets for Taylor Hall at UT Austin during the periods from 6/22/1996-6/22/1997 and 7/17/1997-7/17/1998.
- One-year of hot water data for the Geology Building at UT Austin during the periods from 2/1/1996-2/1/1997.

Table 1 summarizes a comparison of relative error between different model errors for filling missing cooling and heating data gaps based on the average monthly statistical errors for the cooling and heating datasets listed above. This comparison of relative error is based on the error from the polynomial model with temperature as an independent variable and with HOD as an independent variable. The model 2 superior to model 1 if relative error is over than zero. The relative error quantity comparison presents that the error statistics for the HOD models, in the great majority of individual comparisons, are lower than for the temperature-base models.

2. The Comparison of Four Kinds of Different Methods
   (1). Hourly Average Errors Based on Short Period Datasets

Table 1 Comparison of Relative Error for HOD vs Temp. Models (>500,000 Data Gaps)

Relative Error Formula	Gap	Cooling Error				Heating Error			
		MBE (%)	RMSE (%)	CV-SAE (%)	SAE (%)	MBE (%)	RMSE (%)	CV-SAE (%)	SAE (%)
$\frac{E_{T,poly}-E_{HOD,poly}}{E_{HOD,poly}}\times 100\%$	1	165	69	62	86	-25	-13	-6	-14
	2	154	50	72	82	57	3	6	1
	3	36	51	75	82	15	8	10	4
	4	354	59	77	85	73	14	14	12
	5	191	55	77	82	148	16	16	12
	6	92	54	77	81	312	19	18	14

For the comparison of different methods for filling in data gaps, the predicted yearly accuracy (one-year dataset) of each model is important. However, it is also valuable to compare the different models using the same short hourly datasets to show the behavior of the models in filling a specific gap. The development of the different models for the gap treated is based on the same measured short dataset. The polynomial and regression models used 12 hourly measured data points on each side of data gap while the linear interpolation development is based on the 1 measured datum on each side of data gap and the Lagrange interpolation model development is based on the 4 measured data points on each side of the data gap.

Figures 4-5 show an example of the results when the different models are used to fill a gap consisting of 6 consecutive hourly peak cooling values of the Zachry Building (21:00 on Aug. 9, 1998-2:00 on Aug. 11, 1998). Figure 4 shows how each model fills the gap with hour-of-day as the independent variable, and Figure 5 shows how each

model fills the gap with temperature as the independent variable. The 30 measured cooling data points (12 measured data on each side of the data gap are used to develop different models, and the 6 consecutive peak values are assumed to be missing data) are also shown in each graph for comparing measured data with the "filled" values. The line with the black rhombus represents the measured chilled water data and other lines represent the predicted missing peak data points, using different models as indicated in the legend. These different models selected are:

- Regression with temperature as the independent variable (Reg-Temp);
- The polynomial with temperature as the independent variable and the model's order is 8, the average optimal polynomial order, Poly-Temp (8);
- Lagrange with HOD as the independent variable (Lag-HOD);
- Linear interpolation with HOD as the independent variable (Linear).

Table 2 summarizes the accuracy of the different models for filling in 6 consecutive missing peak data points (21:00 on Aug. 9, 1998-2:00 on Aug. 11, 1998) for the Zachry Building. It is easy to see the difference between these predictions using the error parameters. For the linear interpolation model, all the individual errors for filling the data gap have an error of less than 5% and its MBE and RMSE values are –2.36% and 0.09 MMBtu separately, the lowest values among the four models. The polynomial model with HOD as the independent variable, Poly-HOD(8), fills one datapoint (or 17% of the points) with an error of less than 5% and every point (100%) is filled with an error of less than 10% relative error range. The MBE and RMSE values of the polynomial model are separately 6.77% and 0.25 MMBtu, the second lowest values among the four models.

Figures 4-5 also show that the linear interpolation model has the smallest errors (MBE, RMSE, etc.) for filling 6 missing consecutive peak values among the four different models. The polynomial model with HOD as the independent variable (Poly-HOD (8)) is the next best model for filling this data gap. The data patterns show that the Lagrange model and the regression model badly miss the peak values in Figures 4-5. The Lagrange method seems to assume a shorter periodicity in the data, while regression with temperature as the independent variable follows the other side of the hysteresis curve. Hence, both of these models do not come close to the actual data pattern.

The behavior of the models for filling this gap suggests that the regression with temperature as the independent variable and the Lagrange method have serious deficiences for filling gaps as long as 6 hours, and are unlikely to be suitable. On the other hand, linear interpolation and the 8th order polynomial both show considerable promise.

Table 2 Comparison of the Predictive Ability of Different Models Based on Measured Cooling Data on Each Side of Data Gap (9:00 on Aug. 10 - 14:00 on Aug. 10, 1998) for the Zachry Building

Model (Order)	MBE (%)	CV-SAE(%)	RMSE (MMBtu)	Residual (MMBtu)	SAE (MMBtu)	Error<5 %	Error<10 %	Error<15 %
SVR	-21.99	21.99	0.79	-4.66	4.66	0	0	0
Lagrange	-26.91	26.91	1.08	-5.71	5.71	17	17	17
Linear	-2.36	2.86	0.09	-0.50	0.50	100	100	100
Polynomial	6.77	8.19	0.25	1.43	1.43	17	100	100

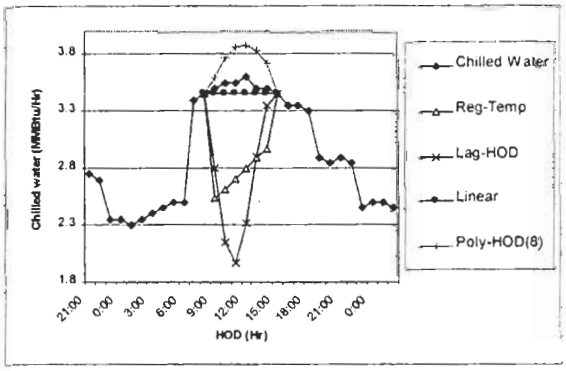

Figure 4 Chilled Water vs HOD grouped by Different Polynomial Models for Predicting 6 Consecutive Peak Cooling Values of the Zachry Building (21:00 on Aug. 9, 1998-2:00 on Aug. 11, 1998)

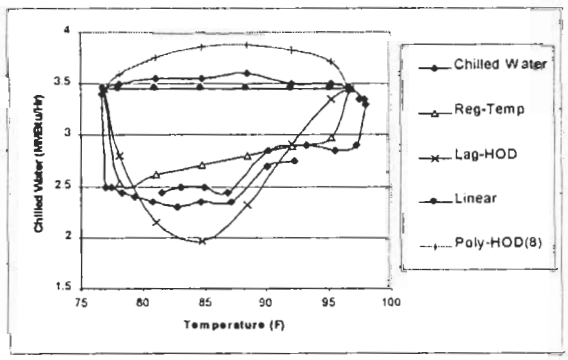

Figure 5 Chilled Water vs Temperature grouped by Different Polynomial Models for Predicting 6 Consecutive Peak Cooling Values of the Zachry Building (21:00 on Aug. 9, 1998-2:00 on Aug. 11, 1998)

(2). The Comparison of Average Errors Based on Different One-year Datasets

Several models were applied to different one-year datasets that were measured in selected LoanSTAR buildings. The main objective of this analytical phase was to confirm the error range when rehabilitating missing hourly data with different interpolation models. The energy types, building names and data periods of the measured datasets for different comparisons are as follows:

- The cooling data of the Zachry Building (9/14/1989-9/14/1990 and 12/20/1991-12/20/1992);
- The heating data of the Zachry Building (9/14/1989-9/14/1990 and 1997);
- The cooling data of the Main Building (4/6/1993-4/6/1994);
- The cooling data of the Geology Building (2/1/1996-2/1/1997);
- The heating data of the Geology Building (2/1/1996-2/1/1997);
- The cooling data of Taylor Hall from 6/22/1996-6/22/1997;
- The heating data of Taylor Hall from 6/22/1996-6/22/1997;
- The cooling data of Taylor Hall from 7/17/1997-7/17/1998;
- The heating data of Taylor Hall from 7/17/1997-7/17/1998.

Table 3 shows a comparison of relative error between different model errors for filling missing energy gaps, which is based on the monthly average statistical parameters from the polynomial model. The RMSE and SAE relative error ranges between the linear interpolation error and the polynomial error are separately around –2.36%~-12.41% and –4.32%~-13.94% for filling 1-6 missing cooling data points, and around -2.28%~12.12% and -3.26%~12.40% for filling 1-6 missing heating data points. The positive values in RMSE and SAE columns and negative values in Error % columns mean the polynomial model is better than the comparison model and vice versa. These RMSE and SAE relative error quantity comparison presents that the linear interpolation model is better than the polynomial model for filling 1-6 missing cooling data points, and the polynomial model is better than the linear interpolation model for filling 1-6 missing heating data points. For the Lagrange model and SVR model, their relative error ranges are far larger than ones of the linear interpolation model. With the increase (from 1 to 6) of data gap number, the Lagrange model become more worse for filling energy use data gap.

Table 3 Comparison of Relative Error between Different Model Errors for Filling Missing Energy Data Gap

Data Type		Cooling Relative Error					Heating Relative Error				
	Gap	RMSE	SAE	Error <5%	Error <10%	Error <15%	RMSE	SAE	Error <5%	Error <10%	Error <15%
$\frac{Elinear - Epoly}{Epoly} \times 100\%$	1	-2.36	-4.32	-1.35	-1.50	-1.00	12.12	12.40	3.54	-0.03	-2.16
	2	-4.94	-7.68	1.13	-0.74	-0.77	11.79	11.96	4.28	0.59	-1.65
	3	-6.79	-10.44	2.69	0.02	-0.12	9.49	9.34	5.89	2.08	0.21
	4	-8.41	-11.58	4.21	1.02	0.14	6.50	5.78	7.67	3.32	1.39
	5	-10.36	-13.07	5.48	1.49	0.29	1.96	0.97	9.30	4.72	2.76
	6	-12.41	-13.94	7.23	2.14	1.19	-2.48	-3.26	11.31	6.87	4.84
$\frac{EReg-Epoly}{Epoly} \times 100\%$	1	110.01	140.53	-36.70	-23.13	-14.30	37.37	42.12	-48.02	-32.02	-18.87
	2	101.39	131.65	-37.58	-24.70	-15.97	42.07	48.75	-48.26	-33.24	-20.95
	3	91.83	119.03	-37.35	-25.45	-16.95	42.49	50.13	-47.99	-34.17	-21.87
	4	82.23	107.45	-36.71	-25.86	-17.65	41.68	49.50	-46.64	-33.98	-22.47
	5	72.36	95.75	-35.67	-25.56	-17.99	35.05	42.52	-44.64	-33.07	-21.75
	6	61.63	83.97	-34.15	-25.03	-17.83	28.10	35.59	-42.04	-31.30	-20.82
$\frac{ELag-Epoly}{Epoly} \times 100\%$	1	25.45	31.07	-12.87	-7.59	-5.01	52.73	55.30	-16.79	-11.16	-10.14
	2	47.85	51.69	-20.05	-12.52	-8.58	91.67	93.43	-29.10	-20.02	-16.69
	3	79.40	82.31	-30.03	-19.29	-13.71	152.17	155.96	-41.42	-31.45	-25.66
	4	119.66	119.07	-39.97	-27.18	-19.83	220.65	221.93	-50.64	-41.61	-34.74
	5	165.93	161.46	-48.41	-35.09	-26.78	295.02	291.04	-57.93	-49.89	-42.96
	6	216.96	207.79	-55.35	-42.79	-33.85	385.45	373.29	-63.19	-56.41	-50.13

The average statistical parameters for filling gaps in dry-bulb temperature data and dew-point temperature data based on the averages of 5 years of monthly values are also calculated. These average MBE, RMSE, SAE and percent of points with less than the specified errors are averaged from the following monthly statistical parameters (errors) obtained when using the Polynomial, Lagrange, regression and linear interpolation methods to fill in 1-6 hour data gaps:

- The temperature of the LoanSTAR Houston weather data during 1995;
- The temperature and dew point of the NWS weather data in Minneapolis from 4/1/1996-4/1/1997;
- The temperature and dew point of the NWS weather data in College Station during 1998;
- The temperature and dew point of the NWS weather data in Washington DC during 1997;
- The temperature and dew point of the NWS weather data in EL Paso during 1997.

Table 4 summarizes a comparison of relative error between different model errors for filling missing weather data gaps. This comparison of relative error is based on the monthly average statistical parameters of 5-year temperature data and 4-year dew point data from the polynomial model. The temperature and dew point temperature RMSE relative error ranges between the linear interpolation error and the polynomial error are separately around –5.35 % ~ +0.24 % for filling 1-6 missing temperature and –34.32 % ~ -12.24 % for filling 1-6 missing dew point temperature data points. The negative values in RMSE and SAE columns and positive values in Error % columns mean the polynomial model is not better than the comparison model. The relative error quantity comparison presents that the linear interpolation model is not only better than the polynomial model for filling 1-6 missing temperature data points, but also better than the polynomial interpolation model for filling 1-6 missing dew point temperature data points. The relative errors in Table 4 also presents that the Lagrange model and SVR model are not better than the polynomial model for filling missing weather data gaps.

Table 4 Comparison of Relative Error between Different Model Errors for Filling Missing Weather Data Gap

Data Type		Temperature Relative Error					Dew Point Relative Error				
Formula	Gap	RMSE	SAE	Error <5%	Error <10%	Error <15%	RMSE	SAE	Error <5%	Error <10%	Error <15%
$\frac{Elinear-Epoly}{Epoly}\times 100\%$	1	-1.04	-11.47	17.77	5.24	2.14	-12.24	-20.06	24.42	7.36	2.93
	2	-0.10	-8.44	14.76	5.40	2.56	-15.19	-21.47	24.63	9.04	4.02
	3	0.24	-6.15	13.59	5.18	2.78	-18.89	-24.58	32.94	12.74	6.24
	4	-0.29	-4.51	8.46	2.31	1.94	-23.41	-28.30	35.37	16.78	8.84
	5	-2.43	-4.58	10.13	1.94	1.74	-28.34	-33.15	50.32	24.54	14.05
	6	-5.35	-5.41	5.62	-0.43	-0.64	-34.32	-37.98	51.87	30.24	18.75
$\frac{EReg-Epoly}{Epoly}\times 100\%$	1	306.15	444.90	-82.02	-74.23	-65.21	61.10	102.94	-52.39	-36.71	-23.52
	2	268.60	380.86	-81.25	-74.32	-66.41	46.13	80.17	-49.61	-36.15	-24.36
	3	235.49	326.78	-79.82	-74.03	-67.10	33.10	60.81	-45.22	-34.28	-23.89
	4	201.60	277.60	-77.96	-73.08	-66.99	20.51	43.17	-40.41	-31.26	-22.32
	5	168.21	231.84	-75.99	-71.58	-66.25	8.75	26.65	-33.74	-26.99	-19.25
	6	136.84	191.38	-73.52	-69.68	-64.93	-3.39	11.60	-27.51	-21.58	-15.38
$\frac{ELag-Epoly}{Epoly}\times 100\%$	1	14.89	7.30	0.06	-0.11	-0.04	7.52	4.74	-2.20	-0.20	-0.32
	2	35.13	24.12	-5.94	-4.38	-2.54	25.27	20.93	-11.43	-6.12	-3.56
	3	55.26	39.67	-12.33	-10.29	-6.94	46.89	38.78	-20.61	-13.07	-8.77
	4	80.27	58.39	-18.75	-16.59	-12.56	71.26	59.23	-29.09	-22.04	-15.92
	5	107.09	75.74	-24.31	-22.09	-18.37	102.75	78.44	-33.42	-28.15	-22.26
	6	123.39	89.29	-28.41	-26.58	-23.37	120.96	94.72	-38.62	-34.35	-28.85

CONCLUSIONS

The comparison of relative error between different models for filling missing energy and weather gaps shows the difference between models most clearly. The linear interpolation model is the simplest and most convenient method, and is superior for filling missing cooling and heating data and missing weather data. The polynomial is a simple yet powerful approach to interpolate missing weather data as well as missing cooling and heating data and provides slightly better RMSE values than linear interpolation for heating data. The polynomial model with HOD as an independent variable is better than the polynomial model with temperature as an independent variable on filling in 6 missing cooling and heating data. The linear interpolation model is a little better than the polynomial model for filling both missing weather data gaps and the missing cooling data gaps. According to the quantitative average error comparison, the polynomial model and the linear interpolation model are comparable and more accurate than other models. The least accurate is the Lagrange model, particularly for larger data gaps. The regression method (SVR) can not deal with missing weather data due to weather data's pattern.

Based on these findings, the linear interpolation model and the polynomial model with HOD as an independent variable are

recommended for filling 1-6 hour gaps in cooling, heating and weather data.

ACKNOWLEDGEMENTS

This research was funded by the Texas State Energy Conservation Office as part of the LoanSTAR Monitoring and Analysis Program.

REFERENCES

Akbari, H., Heinemeier, K.E., Flora,D., and LeConiac,L. 1988 "Analysis of Commercial Whole-Building 15-minute-interval Electric Load Data", *ASHRAE Transactions*, Vol.94. Part 2, pp. 432-449.

Allen, J.C., 1976, "A Modified Sine Wave Method for Calculating Degree Days," *Journal of Environmental Entomology*, Vol. 5, pp. 388-396.

Bronson, D. J., Hinchey, S. B., Haberl, J. S., O'Neal, D. L. 1992 "A Procedure for Calibrating the DOE-2 Simulation Program to Non-Weather Dependent Measured Loads," *ASHRAE Transactions*, Vol. 93 Part 1, pp. 636 - 652

Claridge, D.E., Haberl, J., Katipamula, S., O'Neal, D., Chen, L., Henneghan, T., Hinchey, S., Kissock, K., and Wang, J. 1990a. "Analysis of the Texas LoanSTAR Data," *Proceedings of the Sixth Annual Symposium on Improving Building Systems in Hot and Humid Climates*, Texas A&M University, College Station, TX, Oct. 9 - 10, pp. 53 - 60.

Claridge, D. E., Haberl, J. S., Katipamula, S., O'Neal, D. L., Ruch, D., Chen, L., Heneghan, T., Hinchey, S., Kissock, J. K., Wang, J. 1990b "Analysis of Texas LoanSTAR Data," *7th Annual Symposium on Improving Building Systems in Hot and Humid Climates 1990,* Fort Worth, TX, October pp. 53 - 60

Dhar, A. 1995, "Development of Fourier Series and Artificial Neural Network Approaches to Model Hourly Energy Use in Commercial Buildings," *Ph.D. Dissertation*, Mechanical Engineering Department, Texas A&M University, May.

Erwin K., 1983, *Advanced Engineering Mathematics*, John Wiley & Sons, New York.

Fels., M., 1986, "Special Issue Devoted to Measured Energy Savings, The Princeton Score Keeping Method (PRISM)," *Energy and Buildings*, Vol. 9, Nos. 1 and 2.

Floyd B. Robert and Braddock D. Roger, 1984, "A Simple Method for Fitting Average Diurnal Temperature Curves," *Journal of Agricultural and Forest Meteorology*, Vol. 32, pp. 107-119.

Haberl,J. S. and Bou Saada, T., 1995, *"The USDOE Forrestal Building Lighting Project: Preliminary Analysis of Electricity Savings,"* ASME/JSME/JSES International Solar Energy Conference, HI, pp. 295-304.

Haberl, J.S., Bronson, J.D., and O'Neal, D.L., 1995, "An Evaluation of the Impact of Using Measured Weather Data versus TMY Weather Data in a DOE-2 Simulation of an Existing Building in Central Texas," *ASHRAE Transactions*, Vol. 106, Part 2, pp. 558-576.

Haberl, J.S. and Claridge, D. E., 1987, "An Expert System for Building Energy Consumption Analysis: Prototype Results," *ASHRAE Transactions,* Vol 93, Part 1, pp. 445-467.

Hittle D.C. and Pedersen C.O., 1981, "Periodic and Stochastic Behavior of Weather Data," *ASHRAE Transactions* Vol. 87, Part 2, pp. 173-194.

Hokoi, S., Matsumoto, M., Kagawa, M., 1990, "Stochastic Models of Solar Radiation and Outdoor Temperature," *ASHRAE Transactions,* Vol. 96 Part 2, pp. 245-252.

Katipamula, S., Reddy, T. A., and Claridge, D. E., 1994, "Development and Application of Regression Models to Predict Cooling Energy Consumption in Large Commercial Buildings," *Solar Engineering 1994-Proceedings of the 1994 ASME-JSME-JSES International Solar Energy Conference*, San Francisco, CA, pp. 307-322.

Katipamula, S., Reddy, T. A., and Claridge, D. E., 1995, "Effect of Time Resolution on Statistical Modeling of Cooling Energy Use in Large Commercial Buildings," *Proceedings of the ASME / JSME/ JSES International Solar Energy Conference*, San Francisco, CA, pp. 309-316.

Kemp, W.P., Burnell, D.G., Everson, D.O. and Thomson, A.J., 1983, "Estimating Missing Daily Maximum and Minimum Temperatures," *Journal of Climate Applied Meteorology*, Vol., 22, pp. 1587-1593.

Kissock, J.K., Reddy, T. A., and Claridge, D. E., 1992, "A Methodology for Identifying Retrofit Energy Savings in Commercial Buildings," *The Proceedings of the Eighth Annual Symposium on Improving Building Systems in Hot and Humid Climates*, Texas A&M University, College Station, TX, October, pp. 234 - 246.

Kissock, J.K., 1993, "A Methodology to Measure Retrofit Energy Savings in Commercial Buildings," *Ph.D. dissertation*, Mechanical Engineering Department, Texas A&M University, College Station, Texas, December.

Leslie, N. P. and Reddy, T. A., 1986, "Regression Based Process Energy Analysis System," *ASHRAE Transactions,* Vol. 92, Part 1, pp 23-34.

McCutchan, Morris H., 1979, "Determining the Diurnal Variation of Surface Temperature in Mountainous Terrain," *Journal of Applied Meteorology*, Vol. 18, pp1224-1229.

Pandit, S.M., and Wu S.M., 1983, *Time Series and System Analysis with Applications,* John Wiley & Sons, New York.

Philips, W.F.,1984, Harmonic Analysis of Climatic Data, *Solar Energy,* Vol. 32(3), pp. 319 - 328.

Reddy, T.A., Saman, F., Claridge, D.E., Haberl, J. , Turner, W., and Chalifoux, T., 1997, "Baselining Methodology for Facility-Level Monthly Energy Use---Part 1: Theoretical Aspects," *ASHRAE Transactions*, Vol. 103 Part 2, pp. 505-517.

Ruch, D. and Claridge, D.E., 1992, "A Four-Parameter Change Point Model for Predicting Energy Consumption in Commercial Buildings," *ASME Journal of Solar Energy Engineering*, Vol. 114, pp. 77-83.

Steven, C.C. and Raymond, P.C., 1996, *Numerical Methods for Engineers*, 2d ed., McGraw-Hill, New York.

Wann. M., Doreen Yen and Harvey J. Gold, 1985, "Evaluation and Calibration of Three Models for Daily Cycle of Air Temperature," *Journal of Agricultural and Forest Meteorology,* Vol. 34, pp. 121-128.