INTERNATIONAL JOURNAL OF ENERGY RESEARCH
Int. J. Energy Res. 2006; **30**:1075–1091
Published online 17 May 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/er.1207

Missing data estimation for 1–6 h gaps in energy use and weather data using different statistical methods

David E. Claridge*,† and Hui Chen

Energy Systems Laboratory, Texas A&M University System, College Station, Texas, U.S.A.

SUMMARY

Analysing hourly energy use to determine retrofit savings or diagnose system problems frequently requires rehabilitation of short periods of missing data. This 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 h 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 1-year data sets to develop different models and fill over 500 000 'pseudo-gaps' 1–6 h in length for each model. These pseudo-gaps are created within each data set by assuming data is missing, then these gaps are filled and the 'filled' values compared with the measured values. Comparisons are made using four statistical parameters: mean bias error (MBE), root mean square error, sum of the absolute errors, and coefficient of variation of the sum of the absolute errors. Comparison based on frequency within specified error limits is also used.

A linear interpolation model or a polynomial model with hour-of-day as the independent variable both fill 1–6 missing hours of cooling data, heating data or weather data, with accuracy clearly superior to the single variable linear regression model and to the Lagrange model. The linear interpolation model is the simplest and most convenient method, and generally showed superior performance to the polynomial model when evaluated using root mean square error, sum of the absolute errors, or frequency of filling within set error limits as criteria. The eighth-order polynomial model using time as the independent variable is a relatively simple, yet powerful approach that provided somewhat superior performance for filling heating data and cooling data if MBE is the criterion as is often the case when evaluating retrofit savings. Likewise, a tenth-order polynomial model provided the best performance when filling dew-point temperature data when MBE is the criterion. It is possible that the results would differ somewhat for other data sets, but the strength of the linear and polynomial models relative to the other models evaluated seems quite robust. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: filling data gaps; heating data; cooling data; dry-bulb temperature data; dew-point temperature data

*Correspondence to: David E. Claridge, Energy Systems Laboratory, Texas A&M University System, College Station, Texas, U.S.A.
†E-mail: dclaridge@tamu.edu

Contract/grant sponsor: Texas State Energy Conservation Office

Received 23 June 2004
Revised 11 December 2005
Accepted 2 February 2006

1. INTRODUCTION

Any long-term monitoring effort will have some data records that are missing or bad. These missing or bad records may be due to data processing problems or instrumentation and monitoring hardware problems. The Texas LoanSTAR program monitored energy use data for periods of a year or more from over 200 buildings starting in 1990. This data has been used in conjunction with hourly National Weather Service data to determine retrofit savings and as an aid in diagnosing operating problems in the buildings. About 1% of the weather records are missing (Chen, 1999) and about 2% of the energy records are missing (Haberl *et al*., 1998). However, since daily totals and daily average values are often used for savings determination, a single missing record in a day requires the missing value to be estimated, or the entire day of data to be discarded. Analysis of the missing data showed that all of the missing NWS data and about 60% of the missing energy data was in gaps 1–6 h in length (Chen, 1999).

Three different investigators have reported efforts to fill hourly weather data used for energy simulation. Colliver *et al*. (1995) investigated the use of linear, third-order polynomial, and cubic spline interpolation techniques to obtain 24 hourly readings per day from 3 h data and found that linear interpolation was the best for filling the dew-point temperature gaps and the cubic spline technique provided better results for dry-bulb temperature data. Developing the Typical Meteorological Year data sets used in energy simulation required filling some missing weather data. Gaps of up to 5 h were filled by linear interpolation, except for relative humidity, which was calculated based on measured or filled dry-bulb and dew-point temperature data. Gaps of length 6–47 h were filled by using data from adjacent days for identical hours and then by adjusting the data so that there were no abrupt changes in data values between the filled and measured data (Marion and Urban, 1995). Haberl *et al*. (1995) reported that the DOE-2 weather packer uses linear interpolation to fill weather gaps of less than 24 h.

Numerous other investigators have used a variety of techniques to fill other forms of missing data. Kemp *et al*. (1983) used a linear model and weighted regression to calculate missing daily temperature data within stations in northern and central Idaho. Baker *et al*. (1988) used a linear model to generate hourly temperature data from the daily highs and lows. Acock and Pachepsky (2000) used data from adjacent days to fill missing data maximum and minimum daily temperatures using the so-called group method of data handling. Schneider (2001) used a regularized expectation maximization algorithm to impute missing values of mean July temperatures where spacially adjacent values were present. Others who have investigated techniques for filling missing non-weather data include Beckers and Rixen (2003), Farhangfar *et al*. (2004), Latini *et al*. (2001), Junninen *et al*. (2004), Smith *et al*. (2003), and Sprott (2004).

It is significant to note that while many other investigators have examined the use of techniques for interpolating weather data, only Baltazar and Claridge (2002) have examined techniques for interpolating building energy use data. They examined the use of cubic spline and Fourier series techniques for filling 1–6 h gaps in cooling and heating data series.

This paper evaluates the use of three methods that have not been previously used for filling data gaps of 1–6 h in data sets of dry-bulb and dew-point temperature and commercial building heating and cooling energy use. The methods examined are single variable regression, polynomial 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 with one another and with simple linear interpolation.

2. METHODOLOGY

Five nearly complete 1-year data sets of dry-bulb temperature, dew-point temperature, and heating data and six 1-year sets of cooling data were used to evaluate the different gap filling techniques. Thousands of artificial data gaps, which will be called pseudo-gaps hereafter, were created within each data set and the values estimated by each gap filling technique within each pseudo-gap were compared with the measured values to evaluate the techniques.

Each interpolation model is evaluated for filling data gaps of 1–6 consecutive hours. The gaps evaluated are created by creating a pseudo-gap of a particular length (e.g. 6 h) starting with the 13th hour of a 1-year data set, filling the missing data, and evaluating the errors; the second pseudo-gap created begins with the 14th hour of the data set, the gap is filled, evaluated, etc. until all possible pseudo-gaps in the data set have been created and evaluated The first pseudo-gap starts with the 13th hour of the data set and the last pseudo-gap ends with the 13th hour from the end of the data set since up to 12 h of data on each side of the gap are required by the models used to fill the pseudo-gaps. All pseudo-gaps that can be created in each data set are evaluated, so the maximum number of pseudo-gaps that are created in a complete 8760 h data set varies from 8731 (for 6-h gaps) to 8736 (for 1-h gaps). The number of pseudo-gaps created is reduced by the presence of some real gaps in the data sets used.

The single variable regression and polynomial models use the 12 data points on each side of the pseudo-missing data (24 total points) to create a model and fill the gap. 12 h was chosen after investigating the accuracy of shorter and longer periods on either side (Chen, 1999). 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.

2.1. Criteria used to evaluate the models

The criteria used in this paper to evaluate models for filling data gaps are model accuracy and model simplicity. Model accuracy is the primary criterion, but if two models have comparable accuracy, the simpler model is preferred.

Model accuracy will be expressed in terms of multiple statistical parameters. Minimizing the mean bias error (MBE) is the most important criterion when the data are used for savings determination. However, the root mean square error (RMSE), sum of the absolute value of the errors (SAE), coefficient of variation of SAE (CV-SAE), Error percent, and Relative Error are also used.

The error % is simply the percent error between a single filled pseudo-gap value and the measured value of that point. Measures of gap filling accuracy presented that are based on error % are the percent of filled points that are within 5, 10, and 15% of the correct heating and cooling values, or % of gaps where SAE is within 1, 2 and 3°F of the correct values.

Most of the measures above will be presented as 'average monthly' values. These values are determined as follows, using RMSE as an example. The average value of the RMSE for each month is first calculated for all pseudo-gaps that have been filled during each month in the data set from a particular building or weather station. Then the average of these monthly values is computed for all data sets treated in that specific case.

There are so many individual comparisons to consider that an additional measure has been adopted to simplify the final comparisons. Relative error (RE) has been defined using the

normal definition RE = 100%(Err_2 –Err_1)/ Err_1 where Err_1 and Err_2 are the monthly average values of the errors of models 1 and 2, respectively.

Hence a positive value of the relative error indicates that model 1 is superior to model 2 in this comparison (and vice versa) whenever a small value of the quantity compared is desirable. A negative value indicates that model 1 is superior to model 2 (and vice versa) whenever a large value of the quantity compared is desirable.

2.2. Models investigated

2.2.1. Linear interpolation model. The literature reviewed has heavily used linear interpolation with apparent success, so this approach was included among the techniques to be investigated. The linear interpolation model adopted in this paper is the normal model of the form:

$$f_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) \tag{1}$$

The independent variable, x, was considered to be the time, which was at 1 h intervals in all comparisons in this paper.

2.2.2. Lagrange interpolation model. It is often convenient or possible to use Lagrange interpolation at both equal and unequal intervals (Steven and Raymond, 1996; Erwin, 1983). The Lagrange interpolating polynomial can be represented concisely as

$$p_n(x) = \sum_{j=0}^{n} f(x_i) P_j(x) \tag{2}$$

where

$$P_j(x_k) = \left\{ \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x - x_i}, \quad k = j, 0, k \neq j \right\} \tag{3}$$

The independent variable, x, was considered to be the time, which was at 1 h intervals in all comparisons in this paper.

2.2.3. Polynomial model. A one variable polynomial model is defined as

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m + e \tag{4}$$

where y represents the dependent variable and x the independent variable. The largest 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 lower powers of the independent variable. The least squares method is used to estimate values of the parameters $a_0, a_1, a_2, \ldots, a_m$ that minimize the sum of the squared differences between the actual y values and the values, $\hat{y}_i$ predicted by the equation (Steven and Raymond, 1996; Erwin, 1983).

The independent variable, x, was considered to be either the time, at 1 h intervals, or the ambient temperature, in all comparisons in this paper. It was also necessary to investigate the preferred number of points on either side of the gap to use for gap filling and the optimum order of the polynomial to be used.

2.2.4. Single variable regression model. Building heating and cooling consumption is generally considered to correlate with ambient temperature more closely than with any other variable. While a variety of regression models have been used to model long-term energy use data, the simple two parameter regression model was investigated for use in filling data gaps of 6 h or less with outside air dry-bulb temperature as the only regression variable. The functional form of this model is

$$E = B_0 + B_1 T \tag{5}$$

B_0 is the energy consumption at the intercept $T = 0$ and B_1 is the temperature slope.

2.3. Data sets analysed

2.3.1. Gap length analysis. This study began by examining multi-year hourly data from 87 data channels derived from the National Weather Service (NWS) and from the LoanSTAR database. Data examined included 27 temperature, 20 relative humidity, 14 dew point, eight cooling, eight whole building electricity, seven heating, and three air handler electricity channels. These were multi-year records that included over 300 channel-years of data.

About 2% of the data were missing in both the LoanSTAR data and the NWS data acquired by the LoanSTAR program. The frequency of the LoanSTAR energy use gaps is far lower than the frequency of the LoanSTAR and NWS weather data gaps, but there are more long gaps in the energy use data. Data gaps 1–6 h in length cover almost all missing weather data and the majority of the missing energy use data.

For weather data from both the NWS and LoanSTAR, there are more 1-h data gaps than any other length. The frequency of data gaps with 2 or 3 consecutive missing data hours is far lower than the frequency of gaps with 1 h of missing data. For example, the NWS temperature data from 1 January 1992 to 31 August 1997 for College Station, Texas, was examined. This analysis found that data gaps with 1-h duration account for 1.9% of the total hourly observations; data gaps of 2 h account for 0.18% and data gaps of 3 h for 0.06%.

Analysis of the data sets used found that 1–6 h data gaps covered all missing NWS temperature and dew-point data, 50–70% of total missing LoanSTAR temperature and humidity data, and 50–70% of total missing LoanSTAR energy use data that included data for cooling, heating, motor control centres and whole building electricity use. Hence it was decided to examine methods for filling data gaps of 1–6 h in length in the study reported in this paper.

2.3.2. Data sets used for comparing gap filling models. The first 11 1-year data sets shown in Table I were used for analysing different gap filling models for heating and cooling data unless otherwise noted in the text. The last five data sets shown in Table I were used for analysing the gap-filling techniques for dry-bulb and dew-point temperature data.

3. COMPARISON OF DIFFERENT MODELS

This section compares the interpolation accuracy of all selected models using the measured data sets listed in the previous section.

Table I. Data sets used for analysing gap filling methods for heating and cooling and weather data.

Data type	Building or location	Data period
Cooling	Zachry	9/14/1989–9/14/1990
	Zachry	12/20/1991–12/20/1992
	Main	4/6/1993–4/6/1994
	Geology	2/1/1996–2/1/1997
	Taylor Hall	6/22/1996–6/22/1997
	Taylor Hall	7/17/1997–7/17/1998
Heating	Zachry	9/14/1989–9/14/1990
	Zachry	1/1/1997–12/31/1997
	Geology	2/1/1996–2/1/1997
	Taylor Hall	6/22/1996–6/22/1997
	Taylor Hall	7/17/1997–7/17/1998
Dry-bulb and dew-point temperatures	Minneapolis NWS	4/1/1996–4/1/1997
	College Station, TX NWS	1998
	Washington, DC NWS	1997
	El Paso, TX NWS	1997
Dry-bulb temperature	Houston, TX LoanSTAR	1995

3.1. Determining the optimum order and number of data points for polynomial models

The most suitable number of measured data points for use in model development was also investigated. A short analysis confirmed that use of more than 12 points on either side of a data gap was not appropriate for use with polynomials of order less than 18. The performance of models with five points (order 4), six points (order 5), eight points (order 6), and 10 points (order 8), respectively, on either side of the data gaps, were then compared with the performance of models with 12 points (order 8) on either side of the data gaps (Chen, 1999).

Each set of comparisons was 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 points were then compared with the actual data. Careful examination of the data showed that the models with 12 points or six points on either side generally showed much better statistical performance than the models with 5, 8 or 10 points on either side. Since there was no significant performance difference between six- and 12-point models, it was decided to use 12 points due to the known diurnal cycles prominent in building operation.

After deciding to use 12 points on either side of a data gap to create the polynomial models, the optimum order of the polynomial regression model for heating and cooling data was investigated by first determining the optimum polynomial order for all pseudo-gaps in the January data and in the July data for each of the 11 chilled water and hot water data sets described earlier; this process was repeated for data gaps of 1–6 h lengths. The average value of the optimal order for each gap length was determined and ranged from 9.5 to 8 for the chilled water data and from 9.2 to 7 for the hot water data. This gave choices of 9 or 8 and it was decided to use 8 for both heating and cooling data.

Similar calculations were performed for 8 years of weather data. The best average polynomial order for each gap length for the temperature and dew-point data are almost identical; the range of optimal order for different gap-lengths of temperature and dew-point data is from 11.3 to 9.7.

More details are available in Chen (1999). Polynomials of order 10 were subsequently used for filling data gaps in weather data.

3.2. Comparison of energy use models with time or temperature as the independent variable

When filling data gaps in time series data, time is the natural independent variable for linear interpolation. However, for the heating and cooling use data, dry-bulb temperature would seem to be a more logical variable, since dependence of both on temperature is well known, and gaps in NWS data should be relatively independent of gaps in energy data. Consequently, a detailed comparison of the use of time and temperature as the independent variable for polynomial gap-filling models was conducted. Single variable regression with temperature as the independent variable was also investigated. Clearly, these investigations only apply 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.

Approximately 500 000 gaps were created as described in Section 2 and filled in the 11 1-year heating and cooling data sets noted earlier (except that 12/20/1991–12/20/1992 Zachry heating data was used instead of 1997 data). The errors were summarized and the relative errors are presented in Table II. This comparison of relative error is based on the error from the polynomial model with temperature (T) as the independent variable and with time (t) as the independent variable. Recall that if the relative error is positive, the model with time as the independent variable is superior to the model with temperature as the independent variable. Note that 11 of the 12 MBE comparisons show time to be better than temperature as the independent variable with similar results for the other comparison parameters. Only 1-h heating data gaps show temperature to be a better gap-filling variable than time.

To illustrate the performance of the two approaches for individual data gaps, two 6-h gaps were created around consecutive maximum and minimum values in the 48-h sequence of cooling consumption in Figure 1. This data is taken from the Zachry Building data for 21 July 1992–22 July 1992. From 12:00 to 20:00 and 0:00 to 8:00, there are sequences that are plotted in solid diamonds and solid squares as well as the complete sequence of measured data plotted as open triangles. The middle six hours in each 8-h sequence of diamonds and squares represents the values filled using polynomial temperature and hour-of-day models, respectively. Note that the

Table II. Comparison of relative error for polynomial time (t) vs temperature (T) models (> 500 000 data gaps).

Relative error formula	Gap	Cooling relative error				Heating relative error			
		MBE (%)	RMSE (%)	CV-SAE (%)	SAE (%)	MBE (%)	RMSE (%)	CV-SAE (%)	SAE (%)
$\frac{E_{T,\text{poly}} - E_{t,\text{poly}}}{E_{t,\text{poly}}} \times 100\%$	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

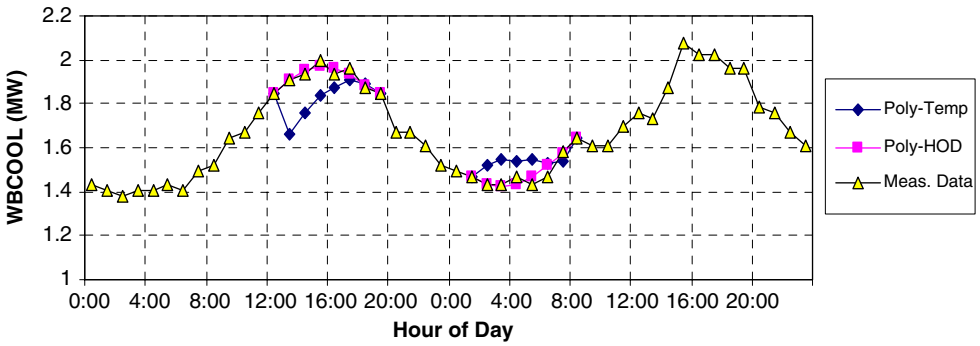

Figure 1. Comparison of time and temperature polynomials used to fill 6-h gaps that include the daily peak consumption and the daily minimum consumption.

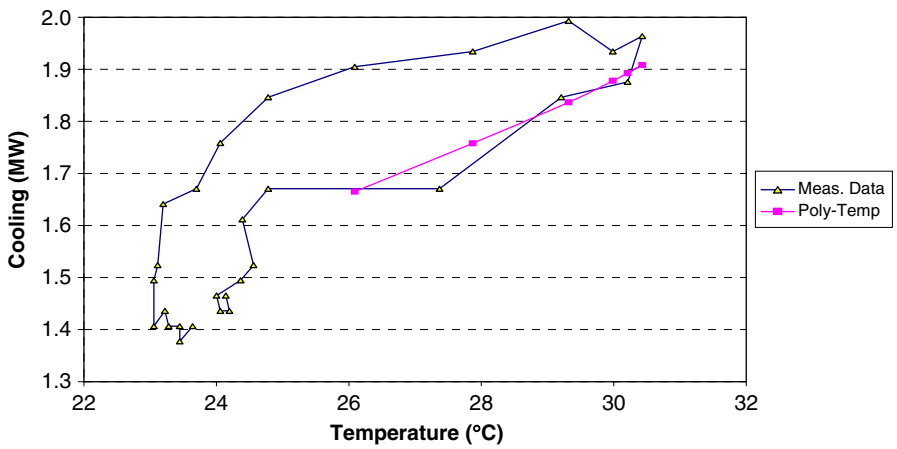

Figure 2. 30 h of cooling data plotted vs ambient temperature with the temperature-dependent polynomial fill of a 6 h pseudo-gap during the daily peak shown in Figure 1.

time-dependent polynomial nicely fills the maximum and minimum values while the temperature-dependent model does poorly. This often occurs because of the hysteresis in short-term heating and cooling data as a function of temperature, primarily due to variations in internal gains. This hysteresis is readily apparent in the 30-h sequence of measured data (triangles) shown in Figure 2.

In Figure 2, it may be noted that the temperature-dependent polynomial values (squares) are at the same temperatures as the measured data points. Looking at the filled values from left to right, the values in the pseudo-gap created are above the interpolated values in every case but one. It can be seen that this is due to the fact that the only remaining values near the temperatures of this mid-day gap are early evening values when cooling consumption was lower than at mid-day. Hence the temperature-dependent polynomial shows substantial error, while the HOD values for this gap in Figure 1 are all within 0.032 MW of the correct value.

Figure 3 compares the use of single variable regression with temperature as the independent variable with polynomial and Lagrangian HOD models and with linear interpolation for a 6 h mid-day gap in cooling consumption on 10 August. This again shows the temperature-dependent model to do a poor job of filling the gap. More extensive analysis showed continued poor performance by the temperature-dependent linear regression model, so it will not be considered further. Figure 3 also shows poor performance by the Lagrangian model, but it will be analyzed further.

3.3. Comparison of linear, polynomial, and Lagrange interpolation models for heating and cooling data

The linear interpolation, polynomial with time as the independent variable, and Lagrangian interpolation models were each applied in the manner described in the 'Models investigated' Section 2. Each model was applied to the 11 different 1-year heating and cooling data sets specified in Section 2.3. Hence each model was used to fill over 500 000 pseudo-gaps of 1–6 h duration, or over 40 000 gaps of each length and data type.

The values shown in Table III for each of the gap lengths from one to six for each of the error measures shown represent the averages of the monthly averages over all the data sets analysed for heating and cooling, respectively. For example, 72 monthly average values of RMSE for filling cooling data gaps of length one are averaged to obtain the value of 0.035 MW shown for Polynomial RMSE. The values shown in the table lines labelled 'Avg.' represent the average of the absolute values for the six gap lengths directly above each value. It may be noted that the MBE for all models and gap lengths is less than 0.014% for cooling data. This suggests that all the models do an impressive job of avoiding systematic bias when filling large numbers of gaps in cooling data. On the other hand, MBE is as large as 0.18% for heating data. This is believed to be a consequence of the much less regular behaviour of heating data generally observed in

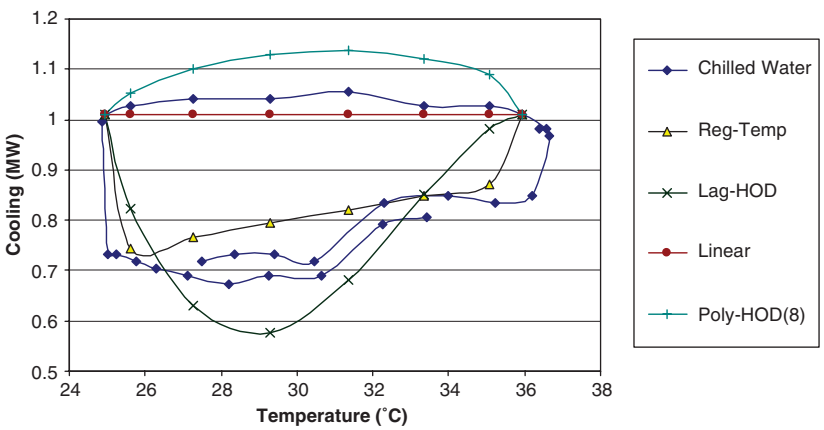

Figure 3. Measured cooling consumption vs temperature from the Zachry Building from 21:00 on 9 August 1998 to 2:00 on 11 August 1998 along with values provided by four different models for filling a 6-h gap near mid-day.

Table III. Averages of monthly average values of error measures for three gap filling models when used to fill all possible pseudo-gaps in at least 5 years of data.

		Cooling comparisons						Heating comparisons					
Model	Data gap	MBE (%)	RMSE (MW)	SAE (MW)	Freq. (CV-SAE <5%)	Freq. (CV-SAE <10%)	Freq. (CV-SAE <15%)	MBE (%)	RMSE (MW)	SAE (MW)	Freq. (CV-SAE <5%)	Freq. (CV-SAE <10%)	Freq. (CV-SAE <15%)
Polynomial	1	−0.001	0.035	0.024	75.5	90.1	94.9	0.020	0.029	0.023	41.8	58.6	66.7
	2	−0.001	0.039	0.052	73.9	89.6	94.8	0.010	0.029	0.045	40.2	57.6	66.7
	3	−0.001	0.042	0.087	71.2	88.4	94.1	0.014	0.030	0.068	38.3	56.6	66.0
	4	0.000	0.047	0.128	68.5	87.0	93.3	−0.005	0.031	0.094	36.0	54.7	64.9
	5	0.000	0.051	0.176	65.7	85.2	92.4	−0.033	0.033	0.127	33.6	52.6	63.1
	6	−0.001	0.056	0.231	62.7	83.2	91.0	−0.047	0.035	0.164	31.3	50.1	61.2
	Avg.	0.001	0.045	0.116	69.6	87.3	93.4	0.021	0.031	0.087	36.9	55.0	64.8
Linear	1	0.000	0.034	0.023	74.5	88.7	93.9	0.021	0.033	0.025	43.3	58.6	65.2
	2	−0.001	0.037	0.048	74.7	88.9	94.0	0.016	0.032	0.050	41.9	58.0	65.6
	3	−0.001	0.040	0.078	73.1	88.4	94.0	0.063	0.033	0.075	40.6	57.8	66.2
	4	−0.002	0.042	0.113	71.4	87.9	93.4	0.083	0.033	0.100	38.7	56.6	65.8
	5	−0.006	0.046	0.153	69.3	86.4	92.6	0.087	0.034	0.128	36.7	55.0	64.8
	6	−0.014	0.049	0.199	67.3	85.0	92.1	0.180	0.035	0.158	34.8	53.5	64.2
	Avg.	0.004	0.041	0.102	71.7	87.6	93.3	0.075	0.033	0.089	39.3	56.6	65.3
Lagrange	1	0.001	0.044	0.031	65.8	83.3	90.1	0.021	0.044	0.035	34.8	52.0	59.9
	2	0.000	0.057	0.079	59.1	78.4	86.6	−0.025	0.055	0.086	28.5	46.1	55.6
	3	0.000	0.076	0.159	49.8	71.4	81.2	−0.011	0.075	0.175	22.5	38.8	49.1
	4	0.001	0.102	0.280	41.1	63.4	74.8	0.023	0.098	0.304	17.8	32.0	42.3
	5	0.003	0.135	0.459	33.9	55.3	67.6	−0.145	0.130	0.496	14.1	26.3	36.0
	6	0.010	0.177	0.712	28.0	47.6	60.2	−0.175	0.172	0.775	11.5	21.8	30.5
	Avg.	0.003	0.099	0.287	46.3	66.6	76.8	0.067	0.096	0.312	21.5	36.2	45.6

buildings where cooling is substantially larger than heating and of the smaller average values of the heating consumption. The RMSE values shown apply to data sets with average CHW consumption of 0.80 MW and HW consumption of 0.32 MW. It is also noteworthy that none of the methods were within 5% of the actual data more than about 75% of the time and some models were within 5% only 20–30% of the time (Lagrange for large gaps).

It is difficult to compare the different models when it is necessary to visually compare many different cells as shown in Table III. Hence Table IV shows the model rankings for each statistical parameter shown in Table III based on the average value of each parameter ('Avg.' in Table III) for all six gap lengths. Note that the model ranked '1' for MBE, RMSE and SAE is the model having the lowest value of these parameters. The model ranked '1' for each of the frequency measures is the model with the highest value rate of occurrence. The rankings shown in Table IV indicate that for both the heating and cooling data, the polynomial model is the best performer based on MBE, while the linear model outranks the polynomial model in seven of the 10 other categories. The Lagrange model is clearly the poorest of these approaches. Looking at the numbers in Table III, it is possible to view both models as providing very good MBE performance, but the polynomial MBE values are less than one-third those of linear interpolation. While linear interpolation outperforms polynomial interpolation using the other measures, the differences are quite small.

3.4. Comparison of linear, polynomial, and Lagrangian interpolation models for weather data

The linear interpolation, polynomial with time as the independent variable, and Lagrangian interpolation models were each applied in the manner described in the 'Models investigated' Section 2 to fill gaps in weather data. Each model was applied to the 5 years of dry-bulb temperature data and 4 years of dew-point temperature data from locations that included hot and dry, hot and humid, and cold climates as specified in Section 2.3. Hence each model was used to fill over 400 000 pseudo-gaps of 1–6 h duration, or over 30 000 gaps of each length and data type, created as described earlier. Single variable regression with temperature as the independent variable was not included in this comparison since dry-bulb temperature is one of the missing variables and the dew-point data is often missing at the same time as the dry-bulb data.

The values shown in Table V for each of the gap lengths from one to six and for each of the error measures shown represent the averages of the monthly average values over all the data sets analysed for dry-bulb and dew-point temperature data. For example, 60 monthly average values of RMSE for filling dry-bulb temperature gaps of length one are averaged to obtain the value of 0.959°C shown for Polynomial RMSE. The values shown in the table lines labelled 'Avg.' represent the average of the absolute values for the six gap lengths directly above each value. We again note that the MBE values are very small for all three models, with the largest MBE 0.006°C and most values 0.001°C or less. However, the other measures are not nearly as good. The RMSE values are all in the range of 0.9–5.4°C while the SAE values cover an even wider range from 0.5 to 17.5°C. However, if we normalized the SAE values by the gap length, the range would be from 0.5 to 2.9°C, indicating that in some cases, the RMSE value is almost doubled by some very large error values. The 'Frequency' values presented represent the percent of the filled gaps where the SAE/n, where SAE is the sum of the absolute errors for the values used to fill an individual gap and n is the gap length, is less than 0.56, 1.11 or 1.67°C. While over 80% of the gaps are filled within 0.56°C when linear interpolation is used for single point gaps,

Table IV. Model rankings for each statistical parameter shown in Table III based on the average value of each parameter ('Avg.' in Table III) for all six gap lengths.

	Cooling comparisons						Heating comparisons					
Model	MBE (%)	RMSE	SAE	Freq. (CV-SAE $<5\%$)	Freq. (CV-SAE $<10\%$)	Freq. (CV-SAE $<15\%$)	MBE (%)	RMSE	SAE	Freq. (CV-SAE $<5\%$)	Freq. (CV-SAE $<10\%$)	Freq. (CV-SAE $<15\%$)
Polynomial	1	2	2	2	2	1	1	1	1	2	2	2
Linear	3	1	1	1	1	2	3	2	2	1	1	1
Lagrange	2	3	3	3	3	3	2	3	3	3	3	3

Table V. Monthly average values of error measures for three gap filling models when used to fill all possible pseudo-gaps in at least 4 years of weather data averaged over at least 48 months.

		Dry-bulb temperature						Dew-point temperature					
Model	Data gap	MBE (°C)	RMSE (°C)	SAE (°C)	Freq. (SAE/n <0.56°C)	Freq. (SAE/n <1.11°C)	Freq. (SAE/n <1.67°C)	MBE (°C)	RMSE (°C)	SAE (°C)	Freq. (SAE/n <0.56°C)	Freq. (SAE/n <1.11°C)	Freq. (SAE/n <1.67°C)
Polynomial	1	0.001	0.959	0.573	68.4	89.3	95.4	0.001	1.216	0.671	64.8	86.6	93.9
	2	0.001	1.122	1.384	60.3	84.4	93.1	0.001	1.401	1.591	57.8	82.2	91.4
	3	0.000	1.299	2.475	52.5	78.7	89.9	0.000	1.601	2.798	50.7	76.7	88.1
	4	0.000	1.512	3.916	45.9	72.6	85.5	0.000	1.830	4.361	44.6	70.8	84.0
	5	0.001	1.767	5.807	40.1	66.0	80.4	0.001	2.091	6.384	38.8	64.5	78.8
	6	0.003	2.068	8.218	35.2	59.8	74.9	0.001	2.419	8.971	34.2	58.4	73.5
	Avg.	0.001	1.454	3.729	50.4	75.1	86.5	0.001	1.759	4.129	48.5	73.2	85.0
Linear	1	0.000	0.949	0.507	80.6	93.9	97.4	0.001	1.067	0.537	80.6	93.0	96.6
	2	0.000	1.121	1.267	69.2	89.0	95.4	0.001	1.188	1.249	72.0	89.6	95.0
	3	0.001	1.302	2.323	59.7	82.8	92.4	0.001	1.298	2.110	67.4	86.5	93.6
	4	0.001	1.507	3.739	49.7	74.3	87.1	0.001	1.402	3.127	60.4	82.7	91.4
	5	0.001	1.724	5.541	44.1	67.3	81.8	0.001	1.498	4.268	58.3	80.3	89.9
	6	0.001	1.957	7.774	37.2	59.6	74.4	0.002	1.589	5.563	51.9	76.0	87.3
	Avg.	0.001	1.427	3.526	56.8	77.8	88.1	0.001	1.341	2.809	65.1	84.7	92.3
Lagrange	1	0.000	1.102	0.614	68.5	89.2	95.3	0.001	1.307	0.703	63.4	86.5	93.6
	2	0.001	1.516	1.718	56.7	80.7	90.7	0.001	1.756	1.924	51.2	77.1	88.1
	3	0.001	2.016	3.457	46.1	70.6	83.6	0.000	2.351	3.883	40.3	66.7	80.4
	4	0.001	2.725	6.203	37.3	60.5	74.7	0.002	3.134	6.943	31.6	55.2	70.6
	5	0.001	3.659	10.206	30.3	51.4	65.6	0.002	4.239	11.392	25.8	46.3	61.3
	6	0.006	4.619	15.557	25.2	43.9	57.4	0.000	5.344	17.468	21.0	38.3	52.3
	Avg.	0.002	2.606	6.292	44.0	66.1	77.9	0.001	3.022	7.052	38.9	61.7	74.4

Table VI. Model rankings for each statistical parameter shown in Table V based on the average value of each parameter ('Avg.' in Table V) for all six gap lengths.

	Dry-bulb temperature					Dew-point temperature				
Model	MBE	RMSE	SAE	Freq. (SAE/n $<0.56^{\circ}$C)	Freq. (SAE/n $<1.11^{\circ}$C)	Freq. (SAE/n $<1.67^{\circ}$C)	MBE	RMSE	SAE	Freq. (SAE/n $<0.56^{\circ}$C)	Freq. (SAE/n $<1.11^{\circ}$C)	Freq. (SAE/n $<1.67^{\circ}$C)
Polynomial	2	2	2	2	2	2	1	2	2	2	2	2
Linear	1	1	1	1	1	1	3	1	1	1	1	1
Lagrange	3	3	3	3	3	3	2	3	3	3	3	3

only 21% of the filled values are within 0.56°C of the true value for 6-point gaps filled by the Lagrange model.

As was done with the heating and cooling data, Table VI shows the model rankings for each statistical parameter shown in Table V based on the average value of each parameter ('Avg.' in Table V) for all six gap lengths. Note that the model ranked '1' for MBE, RMSE and SAE is the model having the lowest value of these parameters. The model ranked '1' for each of the frequency measures is the model with the highest value or rate of occurrence. The rankings shown in Table VI indicate that for dry-bulb temperature data, linear interpolation gave the best performance in every comparison measure, with the polynomial model second in every measure. For dew-point temperature data, the polynomial model offered the best MBE, while linear interpolation was better in every other measure. The Lagrange model is again clearly the poorest of these approaches.

4. CONCLUSIONS

The comparison of the relative error of the polynomial, linear, Lagrangian and single variable regression models evaluated clearly shows that simple linear interpolation and interpolation using a polynomial model with time as the independent variable are clearly superior to single variable regression (with temperature as the independent variable), Lagrangian interpolation, and polynomial interpolation using temperature as the independent variable for purposes of filling 1–6 h gaps in dry-bulb temperature, dew-point temperature, cooling consumption and heating consumption data. This conclusion is based on analysis of nearly 1 000 000 pseudo-gaps created in the data sets used in this study. The linear interpolation model is the simplest and most convenient method, and generally appears superior for filling missing cooling and heating data and missing dry-bulb and dew-point temperature data if statistical criteria other than MBE are used. The eighth-order polynomial model using time as the independent variable is a relatively simple, yet powerful approach that provided somewhat superior performance when filling heating data and cooling data if MBE is the most important criterion. Likewise, a tenth-order polynomial model provided the best performance when filling dew-point temperature data when MBE is the criterion. It is possible that the results would differ somewhat for other data sets, but the strength of the linear and polynomial models relative to the other models evaluated seems quite robust.

ACKNOWLEDGEMENTS

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

REFERENCES

Acock MC, Pachepsky YaA. 2000. Estimating missing weather data for agricultural simulations using group method of data handling. *Journal of Applied Meteorology* **39**(7):1176–1184.

Baker JM, Reicosky DC, Baker DG. 1988. Estimating the time dependence of air temperature using maxima and minima: a comparison of three methods. *Journal of Atmospheric and Oceanic Technology* **5**:736–742.

Baltazar JC, Claridge DE. 2002. Restoration of short periods of missing energy use and weather data using cubic spline and Fourier series approaches: qualitative analysis. *Proceedings of the 13th Symposium on Improving Building Systems in Hot and Humid Climates*, Houston, TX, May 20–23, 213–218.

Beckers JM, Rixen M. 2003. EOF calculations and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and Oceanic Technology* **20**(12):1839–1856.

Bronson DJ, Hinchey SB, Haberl JS, O'Neal DL. 1992. A procedure for calibrating the DOE-2 simulation program to non-weather dependent measured loads. *ASHRAE Transactions* **93**(Part 1):636–652.

Chen H. 1999. Rehabilitating missing energy use and weather data when determining retrofit energy savings in commercial buildings. *M.S. Thesis*, Mechanical Engineering Department, Texas A&M University, December.

Claridge DE, Haberl J, Katipamula S, O'Neal D, Chen L, Henneghan T, Hinchey S, Kissock K, Wang J. 1990. Analysis of the Texas LoanSTAR data. *Proceedings of the Seventh Annual Symposium on Improving Building Systems in Hot and Humid Climates*, Texas A&M University, College Station, TX, October 9–10, 53–60.

Colliver DG, Zhang H, Gates RS, Priddy KT. 1995. Determination of the 1%, 2.5%, and 5% occurrences of extreme dew-point temperatures and mean coincident dry-bulb temperatures. *ASHRAE Transactions* **101**(Part 2):265–286.

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.

Dhar A, Reddy TA, Claridge DE. 1999. A Fourier series model to predict hourly heating and cooling energy use in commercial buildings with outdoor temperature as the only weather variable. *Journal of Solar Energy Engineering* (ASME) **121**:47–53.

Erwin K. 1983. *Advanced Engineering Mathematics*. Wiley: New York.

Farhangfar A, Kurgan L, Pedrycz W. 2004. Experimental analysis of methods for imputation of missing values in databases. *Proceedings of SPIE—The International Society for Optical Engineering*, vol. 5421. Intelligent Computing: Theory and Applications II, 172–182.

Fels M. 1986. Special issue devoted to measured energy savings, the Princeton score keeping method (PRISM). *Energy and Buildings* **9**(1 and 2):1–180.

Forrester J, Wepfer W. 1984. Formulation of a load prediction algorithm for a large commercial building. *ASHRAE Transactions* **90**(Part 1):536–551.

Haberl JS, Bou Saada T. 1995. The USDOE forrestal building lighting project: preliminary analysis of electricity savings. *ASME/JSME/JSES International Solar Energy Conference*, HI, 295–304.

Haberl JS, Bronson JD, O'Neal DL. 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* **106**(Part 2):558–576.

Haberl JS, Claridge DE. 1987. An expert system for building energy consumption analysis: prototype results. *ASHRAE Transactions* **93**(Part 1):445–467.

Haberl JS, Thamilseran S, Reddy TA, Claridge DE, O'Neal D, Turner WD. 1998. Baseline calculations for measurement and verification of energy and demand savings in a revolving loan program in Texas. *ASHRAE Transactions* **104**(2):841–858.

Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. 2004. Methods for imputation of missing values in air quality data sets. *Atmospheric Environment* **38**(18):2895–2907.

Katipamula S, Reddy TA, Claridge DE. 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, 307–322.

Katipamula S, Reddy TA, Claridge DE. 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, 309–316.

Kemp WP, Burnell DG, Everson DO, Thomson AJ. 1983. Estimating missing daily maximum and minimum temperatures. *Journal of Climate Applied Meteorology* **22**:1587–1593.

Kissock JK. 1993. A methodology to measure retrofit energy savings in commercial buildings. *Ph.D. Dissertation*, Mechanical Engineering Department, Texas A&M University, College Station, TX, December.

Kissock JK, Reddy TA, Claridge DE. 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, 234–246.

Latini G, Passerini G, Tascini S. 2001. Air quality data-base implementation by using time series statistic filling. *Advances in Air Pollution*, vol. 10. Air Pollution IX, 587–596.

Leslie NP, Reddy TA. 1986. Regression based process energy analysis system. *ASHRAE Transactions* **92**(Part 1):23–34.

Marion W, Urban K. 1995. *User's Manual for TMY2s Derived from the 1961–1990 National Solar Radiation Data Base*. Version 1.0. National Renewable Energy Laboratory: Golden, CO, and National Climatic Data Center: Asheville, NC.

Philips WF. 1984. Harmonic analysis of climatic data. *Solar Energy* **32**:319–328.

Reddy TA, Saman F, Claridge DE, Haberl J, Turner W, Chalifoux T. 1997. Baselining methodology for facility-level monthly energy use—Part 1: theoretical aspects. *ASHRAE Transactions* **103**(Part 2):505–517.

Ruch D, Chen L, Haberl J, Claridge D. 1993. A change-point principle-component analysis (CP/PCA) method for predicting energy usage in commercial buildings: the PCA model. *Journal of Solar Energy Engineering* **115**:77–84.
Ruch D, Claridge DE. 1992. A four-parameter change point model for predicting energy consumption in commercial buildings. *ASME Journal of Solar Energy Engineering* **114**:77–83.
Schneider T. 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* **14**(5):853–871.
Smith BL, Scherer WT, Conklin JH. 2003. Exploring imputation techniques for missing data in transportation management systems. *Transportation Research Record* **1836**(03-2894):132–142.
Sprott JC. 2004. A method for approximating missing data in spatial patterns. *Computers and Graphics* **28**(1):113–117.
Steven CC, Raymond PC. 1996. *Numerical Methods for Engineers* (2nd edn). McGraw-Hill: New York.