WAVELETS, SELF-ORGANIZING MAPS AND ARTIFICIAL NEURAL NETS FOR PREDICTING ENERGY USE AND ESTIMATING UNCERTAINTIES IN ENERGY SAVINGS IN COMMERCIAL BUILDINGS

A Dissertation

by

YAFENG LEI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2009

Major Subject: Mechanical Engineering

WAVELETS, SELF-ORGANIZING MAPS AND ARTIFICIAL NEURAL NETS FOR PREDICTING ENERGY USE AND ESTIMATING UNCERTAINTIES IN ENERGY SAVINGS IN COMMERCIAL BUILDINGS

A Dissertation

by

YAFENG LEI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Kris Subbarao
Committee Members,	David E. Claridge
	Jeff S. Haberl
	Dennis L. O'Neal
Head of Department,	Dennis L. O'Neal

August 2009

Major Subject: Mechanical Engineering

ABSTRACT

Wavelets, Self-organizing Maps and Artificial Neural Nets for Predicting Energy Use and Estimating Uncertainties in Energy Savings in Commercial Buildings.

(August 2009)

Yafeng Lei, B.S., Tianjin University; M.S., Texas A&M University

Chair of Advisory Committee: Dr. Kris Subbarao

This dissertation develops a "neighborhood" based neural network model utilizing wavelet analysis and Self-organizing Map (SOM) to predict building baseline energy use. Wavelet analysis was used for feature extraction of the daily weather profiles. The resulting few significant wavelet coefficients represent not only average but also variation of the weather components. A SOM is used for clustering and projecting high-dimensional data into usually a one or two dimensional map to reveal the data structure which is not clear by visual inspection. In this study, neighborhoods that contain days with similar meteorological conditions are classified by a SOM using significant wavelet coefficients; a baseline model is then developed for each neighborhood. In each neighborhood, modeling is more robust without unnecessary compromises that occur in global predictor regression models.

This method was applied to the Energy Predictor Shootout II dataset and compared with the winning entries for hourly energy use predictions. A comparison

between the “neighborhood” based linear regression model and the change-point model for daily energy use prediction was also performed.

We also studied the application of the non-parametric nearest neighborhood points approach in determining the uncertainty of energy use prediction. The uncertainty from “local” system behavior rather than from global statistical indices such as root mean square error and other measures is shown to be more realistic and credible than the statistical approaches currently used.

In general, a baseline model developed by local system behavior is more reliable than a global baseline model. The “neighborhood” based neural network model was found to predict building baseline energy use more accurately and achieve more reliable estimation of energy savings as well as the associated uncertainties in energy savings from building retrofits.

DEDICATION

To my family

ACKNOWLEDGMENTS

I would like to express my gratitude to all those who have given me support towards my degree. I gratefully acknowledge Dr. Kris Subbarao for his great guidance and support throughout my research work. I thank Dr. David Claridge for his patience, encouragement and invaluable advice. Special thanks to Dr. Jeff Haberl for his comments and material to improve the content of this study. He broadened my vision in research. Thanks to Dr. Dennis O'Neal for his comments and committee service while he was so busy serving the whole department, Dr. T. Agami Reddy for his support and cooperation in the research, and Dr. Charles Culp for invaluable discussion of my research work. I also thank Ms. Sherrie Hughes and Diane McCormick for their time helping me finish this work. Finally, I would like to express special thanks to my parents and my wife for their endless support and love during the course of my graduate studies.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF FIGURES

Page

Page

Page

LIST OF TABLES

Page

Page

NOMENCLATURE

a_j	Approximation coefficients at level j
ANN	Artificial neural network
BMU	Best matching unit
CV	Coefficient of variations
d_j	Detail coefficients at level j
$dist$	Euclidean distance
DWT	Discrete wavelet transform
E	Building energy use
$E_{error,j}$	Model residual for day j
$E_{meas,j}$	Measured energy use for day j
$E_{\text{model},j}$	Modeled energy use for day j
E_{int}	Internal electrical loads
$E_{savings,j}$	Energy savings for post-retrofit day j
$E_{pre,model,j}$	Modeled baseline energy use for day j
$E_{post,measured,j}$	Measured energy use for day j
h	Filters
I_{sol}	Solar radiation
MLR	Multivariate linear regression

Nbhd	Neighborhood
p	Number of regressor parameters
RH	Relative humidity
$\mathrm{Res}T_{dew}$	Residual dew point temperature
RMSE	Root mean square error
SOM	Self-organizing Map
TMY	Typical meteorological year
T	Outdoor air temperature
$\overline{T}_{out}$	Daily average outdoor air temperature
$\Delta\overline{T}_{dew}$	Daily average effective dew point
ΔT_{dew}	Effective dew point
V_{wind}	Wind velocity
w	Weights
x	Input signal
X	Regressor variables

Subscripts

c	Cooling
dew	Dew point
h	Heating
meas	Measured

out	Outdoor air
post	Post-retrofit
pre	Pre-retrofit

Greek Letters

β	Regression coefficients
Θ	Neighborhood effect function

CHAPTER I

INTRODUCTION

This chapter describes the motivation and objective of the work presented in this dissertation and gives a brief description of the contents of the chapters that are to follow. This section begins by reviewing the function and importance of base-lining models used in industry, followed by a brief introduction of the neighborhood concept and some key analysis tools used in the research.

Motivation

In order to create a cleaner environment, decrease greenhouse gas emission, possibly temper climate change and secure sustainable development, many nations are devoted to developing renewable energy to substitute energy generated from fossil fuels, and at the same time developing new technologies to improve energy efficiency in every element of energy production and end use.

In the United States, the renewable energy portfolio standard has been created in many states which require the electricity provider to obtain a minimum percentage of their power from renewable energy resources by a certain date. The purchased renewable energy are quantified and certified by Renewable Energy Certificates, also known as RECs or Green Tags (Radar and Norgaard 1996). Certificates represent the contractual right to claim the environmental and other attributes associated with electricity generated

This dissertation follows the style of *HVAC&R Research*.

from a renewable energy facility. RECs are traded independently of the energy production. Green Tag purchases are mandated by State Renewable Portfolio Standards (RPS) in some states (EPA 2008).

Similar to Green Tags associated with renewable energy, Energy Efficiency Certificates, also known as White Tag Certificates, are granted to utilities or facility owners which represent the amount of energy conserved through the implementation of energy conservation measures (ECMs). To date, several states have had Energy Efficiency Portfolio Standards (EPS) in place (EPA 2008). In this way, energy savings are monetized by White Tags and can be treated as a commodity through the energy efficiency credit trading market.

For existing building energy conservation retrofits, certification requires the approval of an M&V plan which requires establishing a baseline energy use model to calculate energy savings after implementation of ECMs.

Monitoring and verification (M&V) programs such as FEMP (DOE 2000), IPMVP (DOE 2007) and ASHRAE Guideline 14 (ASHRAE 2002) allow building owners, energy service companies (ESCOs), and financiers of building energy efficiency projects to quantify energy conservation measures (ECMs) performance and energy savings. The determination of energy and cost savings from ECMs requires an accurate baseline model to estimate energy use before implementation of ECMs and the uncertainty of savings to be ascertained properly. An accurate baseline model with low uncertainty is more capable of determining whether an energy efficiency project achieves its goal of improving energy efficiency or not, and is helpful to increase

certainty of risk assessment before implementation of ECMs. In addition to the application in M&V programs, reliable baseline models have been used for fault detection and diagnosis (FDD) of building HVAC&R equipment and optimization of building energy use based on the monitored energy performance data.

Statistical baseline models of building energy use have been widely studied. Among these models, inverse modeling (also called data-driven modeling) is a common and important energy use analysis approach. It allows identification of inherent as-built energy performance based on actual available data, and hence provides a simpler and sometimes more accurate predictor than forward models. Inverse models are mostly used as baseline models to quantify reduction in energy use after building retro-commissioning. For example, the change-point (CP) model is one of the inverse models developed by regressing energy use against outdoor air temperature (T_{out}) (Ruch and Claridge 1991; Kissock et al. 1998). The model must identify a change point from which energy use demonstrates different behaviors according to T_{out}.

All these models are global predictor regression models because regressors spanning the whole regression period are used to develop the models. In this context, the current research aims at developing a "neighborhood" based artificial neural network (ANN) model. The neighborhood days are classified by no only the daily average weather components, but also variations of these weather components during a day. This method, utilizing wavelet analysis and a Self-organizing Map (SOM) (Hagan 1996; Kohonen 2001), is a new method to predict building energy use and estimate uncertainty in potential energy savings.

Purpose and Objectives

The purpose of this study is to promote energy conservation retrofits in commercial buildings by developing a more reliable and robust baseline model for building energy use and by applying the neighborhood concept to decrease uncertainty of energy saving estimation. This methodology, used for hourly and daily energy use prediction with low uncertainty, will benefit M&V projects and building fault detection and diagnosis.

The objectives of the dissertation are to: i) develop a daily energy use ANN model for determination of significant wavelet coefficients of daily weather component profiles, ii) classify neighborhoods based on the significant wavelet coefficients and their weights, iii) develop hourly energy use ANN model to predict baseline energy use, iv) compare the neighborhood-based energy use model to other base-lining methods by using the same data sets, and v) conduct an uncertainty analysis using nearest neighboring days concept.

Description of the Following Chapters

This dissertation is presented in nine chapters. The relevant background and the objectives of the topic have been described. Chapter II is literature review of the related research. The methodology of modeling and two important analysis methods, wavelet analysis and SOM, are presented systematically in Chapter III. Determination of significant wavelet coefficients through the development of a daily energy use neural network model is presented in Chapter IV. Chapter V introduces the U-matrix

representation of SOM and its application to neighborhood classification. A method to develop the hourly energy use model for each of the neighborhoods is presented in Chapter VI. Comparisons between the proposed neighborhood based energy use model and other methods for both hourly and daily energy use simulation are described in Chapter VII. Chapter VIII presents an uncertainty analysis using the nearest neighborhood method. A summary of the present work and possible future directions are presented in Chapter IX.

CHAPTER II

LITERATURE REVIEW

In the past decades, researchers have been dedicated to the study of building energy use modeling that are required by successful building energy conservation programs for calculating energy savings. This literature review covers the previous efforts: (i) the DOE and ASHRAE efforts to establish guidelines for developing M&V procedure and reporting energy savings, (ii) to develop methodologies of modeling building energy performance and (iii) the daytyping methods in building energy analysis. The previous work on uncertainty analysis in building energy modeling is also reviewed.

Review of DOE IPMVP & FEMP Guideline and ASHRAE Guideline 14

In order to provide general guidelines for retrofits performed under performance contracts, DOE (1997) released the first edition of International Performance Measurement and Verification Protocol (IPMVP). The latest version, IPMVP 2007 (DOE 2007), is a guidance document that provides a conceptual framework in measuring, computing, and reporting savings achieved by energy or water efficiency projects at facilities. IPMVP provides a framework and definitions that can help practitioners develop M&V plans for their projects. The IPMVP is probably best known for defining four M&V Options for energy efficiency project measurement and verification. They are partially measured retrofit isolation, retrofit isolation, whole facility and calibrated simulation. The FEMP M&V Guideline (DOE 2000) contains specific procedures for

applying concepts originating in the IPMVP. The Guideline represents a specific application of the IPMVP for federal projects. It outlines procedures for determining M&V approaches, evaluating M&V plans and reports, and establishing the basis of payment for energy savings during the contract. These procedures are intended to be fully compatible and consistent with the IPMVP. Compared to the IPMVP, the FEMP Guidelines provide similar background information, but more detail on specific M&V techniques.

In 2002 ASHRAE published Guideline 14-2002 (ASHRAE 2002) to fill a need for a standard set of energy and demand savings calculation procedures (Haberl et al. 2005). The guideline provides three approaches that can be used to measure retrofit savings. The three approaches are the whole building approach, the retrofit isolation approach and the calibrated simulation approach. This guideline is fairly technical document that addresses the analysis, statistics and physical measurement of energy use for determining energy savings. The three approaches presented are closely related to and support the options provided in IPMVP.

The whole building approach in the previous guidelines requires establishment of baseline energy use model to measure energy retrofit savings. For example, ASHRAE Guideline 14-2002 and FEMP M&V Guideline recommend the use of 1, 2 parameter models, change-point model and multivariate models. In M&V programs, daily data based regression models provide satisfied goodness of fit especially for cooling energy use (Katipamula et al. 1995), and hourly data based modeling can provide information

for system monitoring, fault detection and optimization. The next section will review some recognized building energy use simulation models.

Review of Inverse Analysis Methodologies

The currently used methods for analyzing building energy use can be classified into three groups: forward modeling, inverse modeling and hybrid modeling (Rabl ea al. 1986; Rabl, 1988). According to the definition by Rabl, forward modeling is most often used in building design stage for load calculation, HVAC system design and associated design optimization etc. because the system behavior can be predicted before it is physically built by forward modeling. Major government funded energy simulation codes like DOE-2 (LBL, 1993), BLAST (BSL, 1999), and EnergyPlus (UIUC and LBNL, 2005), are in this category. Inverse modeling, which is also known as data-driven modeling, is used to establish an empirical relationship between the energy performance and some variables that affect building energy consumption. Inverse models are usually used as baseline models to predict baseline energy use after retrofitting. The hybrid modeling has the signatures from both the forward and the inverse models. By tuning or calibrating the input variables of an established physical model to match the observed energy use, more reliable predictions can be made. This study focuses on the inverse modeling. Some inverse modeling methods will be reviewed and discussed.

For single-zone buildings that the energy uses are primarily influenced by the envelope, such as residential and small commercial buildings, space-heating energy use increases as outdoor air temperature decrease below a certain balance temperature. The

heating balance temperature is defined as the outdoor air temperature at which the internal heat gain balances heat loss through the building envelope (ASHRAE 2005). At outdoor air temperatures above the balance temperature, no thermal energy is needed for space heating. However, hot water is still required by other end use. Similarly, cooling energy use increases as outdoor air temperature increases above a certain balance temperature. At outdoor air temperature below the balance temperature, no space cooling is necessary but energy (such as electricity) is still needed for other applications. The energy use patterns for single zone buildings are shown in Figure 2.1.

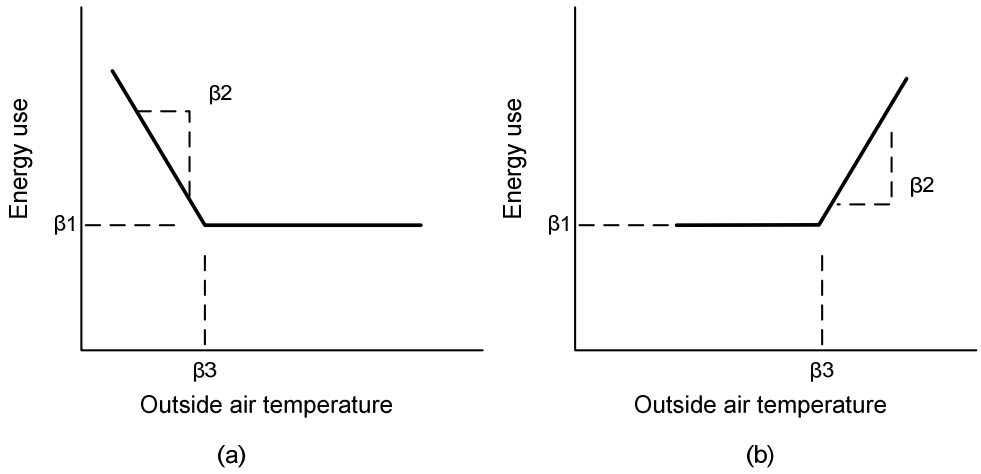


Figure 2.1. Energy use in heating season (a) and energy use in cooling season (b).

A degree-day model establishes a linear correlation between seasonal degree-days computed at a set balance temperature and energy consumption to predict energy use. PRISM model (Fels 1986a; Fels and Goldberg 1986) adopted this method for use in measuring savings through PRInceton Scorekeeping Method. The balance temperature is

determined by finding the best statistical fit between energy consumption and degree-days during an energy use period. The method is widely used in evaluation of resident energy conservation programs, and it also provides adequate statistical fits with commercial building billing data (Eto 1998; Haberl and Vajda 1998; Haberl and Komer 1990; Kissock and Fels 1995; Sonderegger 1998). The PRISM heating-only and cooling-only models are special case of three-parameter (Kissock et al. 1998).

Although the variable-base degree-day method is suitable for residential and single zone commercial building energy use evaluation, it cannot be applied to commercial buildings with simultaneous heating and cooling load (Rabl et al. 1986; Kissock 1993). This is because heating and cooling load in a multizone building vary with outdoor air temperatures, and the energy consumption above or below the balance temperature is not a constant. Therefore, the use of a variable-base degree-day model or three-parameter model will not accurately predict the energy use above or below the balance point. For that case, the four parameter change-point (CP) models (Schrock and Claridge 1989; Ruch and Claridge 1991; Kissock et al. 1998) are superior to variable-base degree-day models to account for nonlinear relationship. Figure 2.2 depicts a 4-parameter cooling and heating CP models.

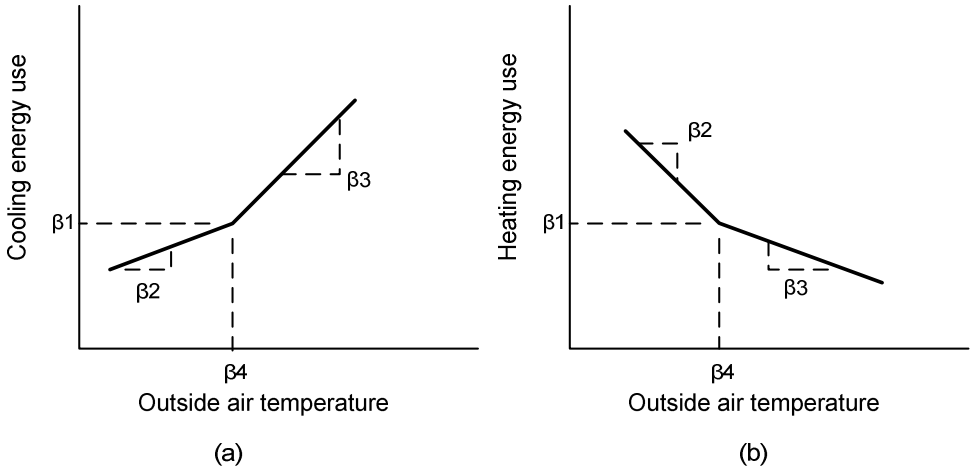


Figure 2.2. 4P cooling and heating CP models.

Single variable regression models use outdoor air temperature as the only regressor because outdoor air temperature is the most important factor that affects energy use in buildings that are dominated by envelope or ventilation load. Practically, weather dependent energy use such as cooling energy use is affected by not only outdoor air temperature, but also dew point temperature, solar radiation, internal gain etc. For example, in commercial buildings, a major portion of the latent load derives from fresh air ventilation. This makes dew point temperature a significant input when establish an energy use model. A number of researchers have studied multivariable regression analysis (Boonyatikarn 1982; Leslie et al. 1986; Mazzucchi 1986; Haberl and Claridge 1987). Fowlkes (1985) proposed an all weather-based parameter regression model for analyzing residential energy use. Multivariable regression showed promise in modeling building energy use but one concern is the determination of which variables should be used to develop the model and how can intercorrelations between independent variables

be removed (MacDonald and Wasserman 1989). The multivariable regression model is a logical extension of single-variate model provided that the choice of variables to be included and their functional forms are based on the engineering principles. Katipamula et al. (1994) developed a multivariable regression model based on engineering principles of the systems used in the building. In the current study, multiple variables will be in the consideration of model input selection.

Regression models are considered reliable retrofit saving models for commercial buildings that have thermostatically controlled HVAC systems (Claridge et al. 1992; Reddy et al. 1994). However, there are some buildings for which change-point linear models do not fit the data adequately especially when the buildings exhibit non-linear behaviors. To improve retrofit saving modeling accurate, the concept of an inverse hourly bin modeling approach (Thamilseran and Haberl 1995) was proposed for those buildings where the regression-based models do not describe the pre-retrofit baseline energy use adequately. This approach can be used to determine weather independent retrofit saving and weather independent retrofit saving. In this model, the data were also separated into four humidity groups to represent different humidity level. The humidity bins were used to account for the fresh air latent load in the system. This approach is an improvement over change-point models because of its ability to handle multiple change-points (Thamilseran and Haberl 1995).

All of the methods discussed so far are steady state method except inverse bin method which utilized the lagged temperature to take into account building thermal mass delay to the response of heating and cooling. In general, steady state inverse models are

used with monthly and daily data containing one or more independent variables. Dynamic inverse models are usually used with hourly or subhourly data because energy use at hourly or subhourly level may significantly affected by building's thermal mass. In the past decades, researchers have been dedicated to the study of dynamic methods (for review, see Rabl 1988; Reddy 1989). A traditional dynamic model requires solving a set of differential equations. Dynamic inverse models based on pure statistical approaches have also been reported such as artificial neural networks (Kreider and Haberl 1994; Kreider and Wang 1991; Miller and Seem 1991). A dynamic method particularly suited for reconciling simulations with data is given by Subbarao (Subbarao 2001). Artificial neural networks are considered heuristic because the neural networks learn by example rather than by following programmed rules. Neural networks have the capability to handle large and complex systems that simple linear regression models are unable to simulate.

ASHRAE organized an open competition in 1993 in order to identify the most accurate method for making hourly energy use predictions based on limited amounts of measured data (Kreider and Haberl 1994). The first place winner employed the Automatic Relevance Determination (ARD) model, a method of automatically detecting relevant input variables based on Bayesian estimation, to develop neural networks that had a single hidden layer with 4-8 tanh units (MacKay 1994). Five of top six winners among more than 150 contestants adopted neural network models to predict hourly building energy use in this contest except one who used piecewise linear regression model. A second predictor shootout contest has been held to evaluate whole-building

energy use baseline models for purpose of measuring savings from energy conservation retrofits (Haberl and Thamilseran 1996). The first place winner ((Dodier and Henze 1996) used a Wald test which is similar to the ARD model to select the only relevant input variables of neural networks.

The combination of neural network with wavelet has also been studied (Dhar 1995). Instead of using most commonly used linear or power transfer functions, wavelet basis functions were used as ANN transfer function. This approach, called Wave-Net approach, can provide better localization characteristics. Moreover, the number of basis functions can be optimized by retaining only the statistically significant smoothing components (known as scaling functions) and detailed components (known as wavelets). This results in more compact network architecture (Dhar 1995). For hourly energy simulation, the current work will use artificial neural network model. Wavelet will also be used in this research but in a totally different manner compared to Dhar's work.

Day-typing Methods

Energy uses in commercial buildings are affected in a major way by two factors: system operation schedule and weather conditions. Classifying the dataset based on operational changes and daily meteorological feature before the model development would be helpful. Katipamula and Haberl (1991) proposed a simple statistical daytyping methodology to identify diurnal load shapes from hourly non-weather dependent loads data. This method sorted the whole dataset into low, high and normal groups and weekday/weekend subgroups by a predetermined standard deviation limit of 24-hour

load profile. The resulting load shapes can be used in a calibrated DOE-2 simulation to represent equipment and occupancy schedules. Akbari (1988) studied the disaggregation of commercial whole-building hourly electrical load into end uses by regressing hourly load against temperature over summer season and winter season. However, all the summer days and winter days are business days only and the method is limited to the cooling season. Thamilseran (1999) studied daytyping method by separation of weekday, weekend and holidays. Duncan's, Duncan-Waller's or Scheffe's multiple comparison tests can be performed to aggregate any daytypes which have means with statistically insignificant differences. To achieve a more accurate day type classification and broader application, Bou-Saada and Haberl (1995) proposed a weather daytyping procedure for disaggregating hourly end-use loads in a building from whole-building hourly data. In this method, three daytypes are divided into temperatures below 7°C, temperatures between 7°C and 24°C, and temperatures above 24°C. The three day types were further divided into weekday and weekend sub-daytypes. Separating weather daytypes by dry bulb temperature only is the simplest way. Hadley (1993) did some improvements by proposing a weather day-typing method to identify distinctive weather day types. In this method, principle components of six meteorological variables, dry-bulb temperature, wet-bulb temperature, extraterrestrial radiation, total global horizontal radiation, clearness index and wind speed were considered for day type classification. Although this method provided additional information about the relationship between climate and the pattern of HVAC system consumption, There are still two concerns. One is that all the variables are daily average data. Daily average weather data may not informative

enough to represent daily weather variation. The other is that each meteorological variable has different influence (or weight) on building energy use. The weights are needed to be considered for accurate day type classification.

The current work will find out the day types by classifying the meteorological days into different neighborhood using wavelet analysis on outdoor air temperature, dew point temperature and solar radiation. Wavelet coefficients represent not only daily average of weather variables but also overall daily weather profile. Weights of these weather variables will be determined for the neighborhood classification. The operational schedules will be considered as input to the energy use neural network model for hourly energy simulation. The daytype routine developed in this research will be an improvement over the previous work.

Discrete Wavelet Transform

Historically, the concept of "wavelets" originated from the study of time-frequency signal analysis, wave propagation, and sampling theory. One of the main reasons for the discovery of wavelets and wavelet transforms is that the Fourier transform cannot be used for analyzing signals in a joint time and frequency domain (Debnath 2002). Wavelet analysis was first introduced by Haar (Haar 1910). In 1982, Morlet developed wavelets as a family of functions constructed by using translation and dilation of a single function for the analysis of non-stationary signals. Wavelet analysis has gradually come to maturity since 1980s with the discovery of orthogonal wavelet basis (Grossman and Morlet 1984; Mallat 1988).

Wavelet analysis is an exciting new method used for data compression, image processing, pattern recognition, computing graphics, and other medical image technology (Addison 2002). For example, the Federal Bureau of Investigation has adopted a wavelet-based image-coding algorithm as the national standard for digitized fingerprint records because wavelet transform provides the ability to represent complicated signals accurately with a relatively small number of bits (Brislawn 1995). Another application is JPEG 2000 wavelet-based image compression standard created by the Joint Photographic Experts Group committee. It was created in 2000 with the intention of superseding their original discrete cosine transform (DCT) based JPEG standard. This new image compression standard has a broad range of functionality, as well as an excellent compression rate (Usevitch 2001; Smith 2003).

Self-organizing Map

The Self-organizing Map was developed by Kohonen in the early 1980s (Kohonen 1981 and 1982). The first application area of the SOM was speech recognition (Kohonen et al. 1984). Other applications of SOM in engineering include identification and monitoring of complex machine and process states, pattern classification and target recognition, and fault diagnosis (Kohonen and Simula 1996; Kohonen 2001).

Uncertainty Analysis

Uncertainty associated with the single-variate linear and multivariate linear regression energy use models can be deduced from rigorous statistical theory (Neter

1989). Goldberg (1982) estimated the uncertainty of model parameters in the PRISM method. A simplified method to estimate the uncertainty of energy savings has been described by Reddy et al. (1998) and Kissock et al. (1998). Reddy and Claridge (2000) suggested that model be evaluated by the ratio of the expected uncertainty in the savings to the total savings. To predict uncertainty of energy use model with autocorrelated residual, Ruch et al. (1999) proposed a hybrid of ordinary lease squares and autoregressive model. This method is based on the assumption that the building energy use is a pure linear function of temperature. The autocorrelation may be caused by time dependent operational changes in the building or by the omission of variables that may influence energy use, such as humidity and solar radiation (Ruch et al. 1999). The energy use model in the current work will consider the operational schedule and multiple variables to reduce residual autocorrelations.

All the previous works determine uncertainty based on the rigorous statistical algorithm that considers all the input and output variables. This method must presume a probability distribution of the output model errors. The direct analytical estimation of the probability distribution of the model error is often impossible; but it can be determined by the quantiles, or prediction interval of the model prediction. This dissertation proposed a robust method to estimate uncertainty of energy use prediction and energy saving prediction. This method is independent of the model structures and requires only model outputs.

Summary

In this chapter, a brief literature review on various measurement and verification protocols, inverse analysis methodologies, daytyping methods and uncertainty analysis was presented. Compared to the previous work, a new daytyping method using wavelet analysis will be proposed and its application in both hourly and daily energy use prediction will be presented in the following chapters. Inspired by this new daytyping concept, the nearest neighboring days method to improve uncertainty estimates in statistical building energy models will be introduced. This method will provide a more realistic, robust and credible way in uncertainty analysis compared to the previous work.

CHAPTER III

OVERVIEW OF THE METHODOLOGY

This chapter provides an overview of the methodology developed in this dissertation to measure building baseline energy use and retrofit savings. It starts from background introduction of discrete wavelet analysis (DWT) and Self-organizing map (SOM) used in the methodology. In the next several chapters, the entire methodology is demonstrated with detailed examples.

Discrete Wavelet Analysis

Introduction of Wavelet Analysis

Wavelet analysis can be viewed as an alternative to Fourier analysis for the purpose of identifying nonlinear behavior of both continuous-time and discrete-time functions. The significant advantage that makes wavelets superior to Fourier analysis for function approximation is their localization characteristics. Just like Fourier analysis that analyzes input signals at different frequency, wavelet analysis analyzes signals at different space levels as well as frequency. This is also called wavelet multiresolution representations. Approximation of a signal in a multiresolution hierarchy is advantageous when signal is nonuniformly distributed in the input space. A high density signal may need higher resolution (high level of space) to be represented and a low density signal may need lower resolution (low level of space) to be represented. Wavelet

transform can be used to analyze frequency features of a signal at different time locations. This makes wavelet transform more powerful than Fourier Transform.

Multirate Processing and Filter Banks

The process of discrete wavelet analysis (DWT) for a signal $x[n]$ is to pass the signal through a series of filters at different levels. Figure 3.1 illustrates the DWT process for an input signal $a_j[n]$ decomposition to calculate wavelet coefficients: approximation coefficients and detail coefficients. Approximation coefficient $a_{j+1}[n]$ at a lower level is determined by passing the signal through a low pass filer h_0 and downsampling of factor 2. Detail coefficient $d_{j+1}[n]$ is determined by passing the signal through a high pass filter h_1 and downsampling of factor 2. It is important that the two filters are related to each other to make alias term zero, and they are known as quadrature mirror filter.

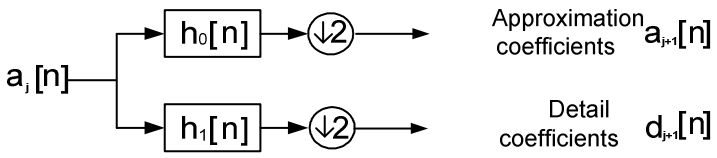


Figure 3.1. Block diagram of filter analysis.

The filtering process is called correlation. Mathematically, correlation followed by downsampling of factor 2 can be combined into one equation:

$$a_{j+1}[n] = \sum_{k=-\infty}^{\infty} a_j[k] h_0[k-2n] \tag{3.1}$$

$$d_{j+1}[n] = \sum_{k=-\infty}^{\infty} a_j[k] h_1[k-2n] \tag{3.2}$$

This decomposition is repeated to further increase the frequency resolution, and the approximation coefficients are decomposed with high and low pass filters and then down-sampled. This is represented as a binary tree with nodes representing a sub-space with different time-frequency localization. The tree demonstrated in Figure 3.2 is known as a filter bank.

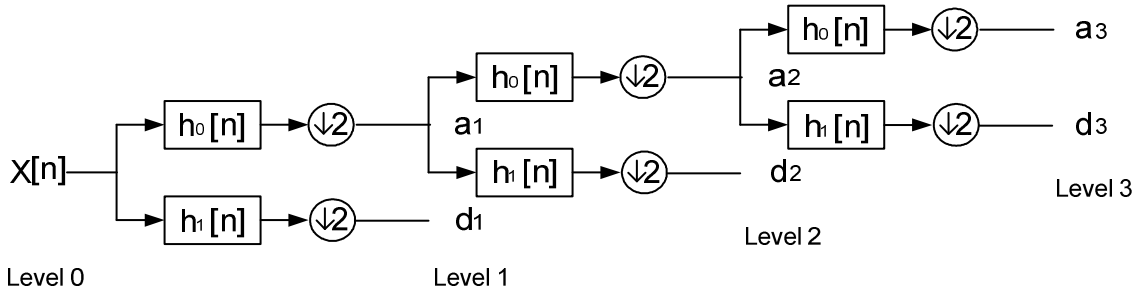


Figure 3.2. Analysis process.

The wavelet used in the current study is Daubechies wavelet Db3. Its corresponding low pass filter h_0 and high pass filter h_1 are the following:

$$h_0 = [0.2352 \;\; 0.5706 \;\; 0.3252 \;\; -0.0955 \;\; -0.0604 \;\; 0.0249]$$

$$h_1 = [0.0249 \;\; 0.0604 \;\; -0.0955 \;\; -0.3252 \;\; 0.5706 \;\; -0.2352]$$

Input signal x can be recovered by upsamping of a factor 2 and filtering. This synthesis process is illustrated in Figure 3.3 h_2 is called the conjugate quadrature filter of h_0.

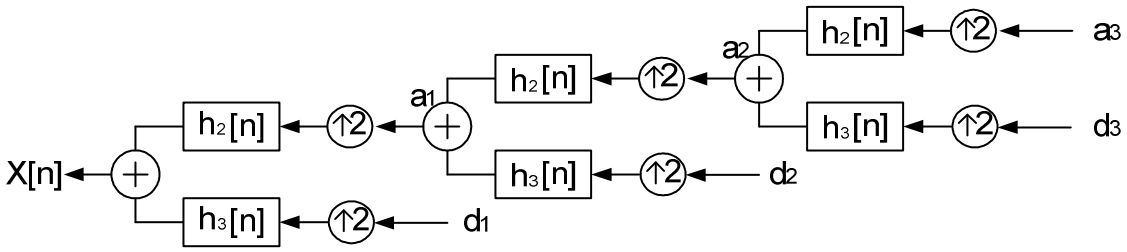


Figure 3.3. Synthesis process.

Filtering in synthesis process is called convolution. Upsampling of factor 2 followed by correlation is expressed as $\sum_{k=-\infty}^{\infty} a_{j+1}[k]h_2[k-2n]$ and $\sum_{k=-\infty}^{\infty} d_{j+1}[k]h_3[k-2n]$. So, approximation coefficient at higher level can be calculated as:

$$a_{j+1}[n] = \sum_{k=-\infty}^{\infty} a_j[k]h_2[n-2k] + \sum_{k=-\infty}^{\infty} d_j[k]h_3[n-2k] \tag{3.3}$$

For Daubechies wavelet Db3, the filters h_2 and h_3 corresponding to synthesis process are as follows:

$$h_2 = [0.0249 \;\; -0.0604 \;\; -0.0955 \;\; 0.3252 \;\; 0.5706 \;\; 0.2352]$$

$$h_3 = [-0.2352 \;\; 0.5706 \;\; -0.3252 \;\; -0.0955 \;\; 0.0604 \;\; 0.0249]$$

Multiresolution of Discrete Wavelet Transform

Discrete decomposition of signal into multi-levels is known as multiresolution. A 5 level multiresolution representation, illustrated in Figure 3.4, shows the procedure of DWT of an input signal with 32 components. j at 0 is the starting level. Different levels represent different frequency characteristics. The amount of levels is determined by the signal density. For example, a signal with 32 components can be decomposed into 5 levels ($32 = 2^5$) according to dyadic grid arrangement. Level 0 is original signal. Level 1 contains d_1 (16 detail coefficients) and a_1 (16 approximation coefficients) after the decomposition of the original signal. a_1 can be decomposed to level 2 which contains d_2 (8 detail coefficients) and a_2 (8 approximation coefficients). This decomposition will continue down to the last level.

Each level represents the input signal by a particular coarseness. From methodological point of view, the highest level of the multiresolution representation should not be an original input signal. It should be a decomposed approximation

coefficient. Common practice is to set the input signal as the approximation coefficients at scale zero if the variation of the signal is not too sharp.

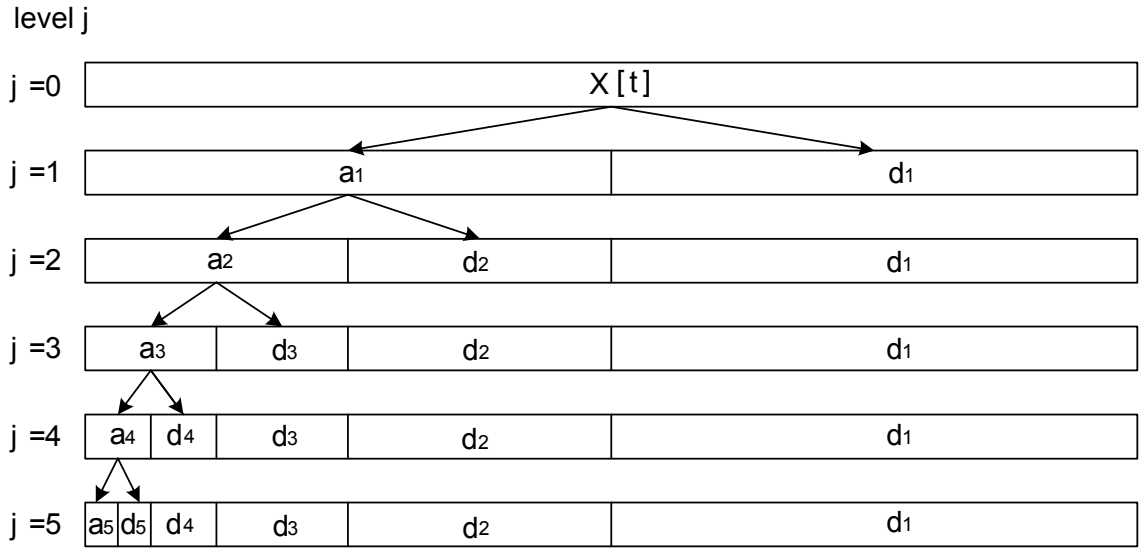


Figure 3.4. Schematic diagram of wavelet decomposition.

In general, the analysis equations decompose an input signal $x[t]$ into approximation coefficient $a_j[k]$ and detail coefficient $d_j[k]$. The synthesis equation builds the signal $x[t]$ from its coefficients $a_j[k]$ and $d_j[k]$.

An Example of Discrete Wavelet Transform

The following example demonstrates how discrete wavelet analysis decomposes an input signal, i.e., a sequence of hourly temperature of a day. In this example, the cubic spline interpolation (Stoer and Bulirsch 1996) applied to interpolate original 24

points to 32 points to agree with dyadic grid arrangement which requires the dimension of input signal is the power of 2. Figure 3.5 shows an example of cubic spline interpolation of T_{out} in 4/4/1990, College Station, TX.

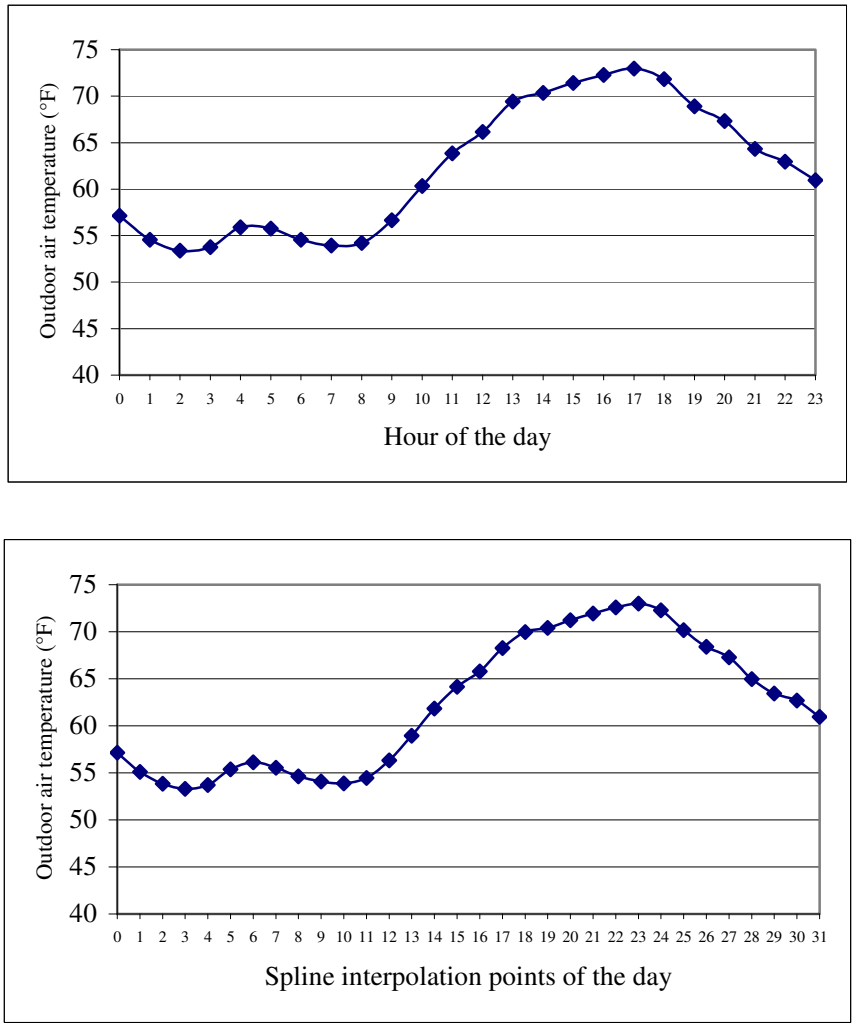


Figure 3.5. Spline interpolation of T_{out}.

Table 3.1 lists the resulting coefficients of DWT on interpolated signal of $x[t]$ using Haar wavelet.

Table 3.1 Wavelet coefficients of daily outdoor air temperature in 4/4/1990 in College Station using Haar wavelet decomposition

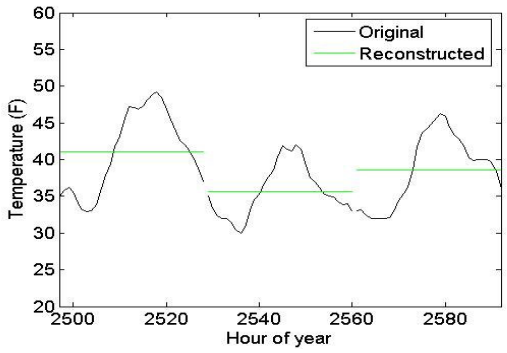
	wavelet coefficients															
d_1	1.45	0.39	-1.2	0.42	0.39	-0.4	-1.9	-1,7	-1.7	-0.3	-0.5	-0.3	1.48	0.81	1.08	1.21
d_2	2.56	-1.3	0.21	-5.4	-3.2	-1.2	3.4	2.4								
d_3	-0.49	-8.6	-5.1	9.22												
d_4	-4.6	8.2														
d_5	-34.5															
a_5	352.1															

Wavelet transform uses little wavelike functions (or filters) to transform the signal under investigation into another representation which presents the signal information in a more useful form. Mathematically speaking, the wavelet transform is a convolution of the wavelet function with the signal to find the approximation coefficient *a* and detail coefficient *d* at a certain level. Once we have the approximation and detail coefficients, the original signal can be reconstructed perfectly using all the coefficients. In this study, some significant coefficients are identified. Scale thresholding is employed to set detail coefficients to be zero for some low levels and keep only approximation

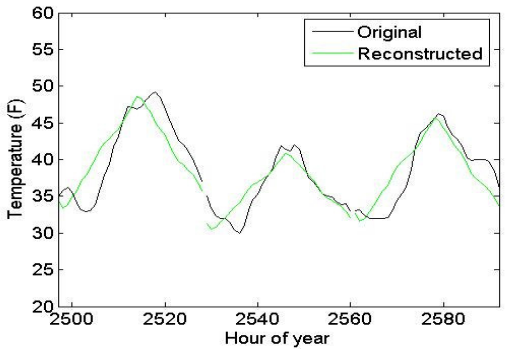
coefficients as well as detail coefficients at high levels. These retained significant coefficients characterize the behavior of the original signal at a compressed form and keep most of the original signal energy with only small distortion. Figure 3.6 is an example of using significant coefficients to represent the original signal. In this example, the original outdoor air temperature profiles of three consecutive days in 1990, College Station, TX, were compared to the reconstructed temperature profiles using first 1, 2, 3 or 4 coefficients. It is very clear that the first coefficient, T_{out} at a_5 is daily average temperature. Reconstruction with T_{out} at a_5 and d_5 captures temperature variation trend during the day. Reconstructions with T_{out} at a_5, d_5, and d_4 can capture more detailed variation in the original data.

Self-organizing Map

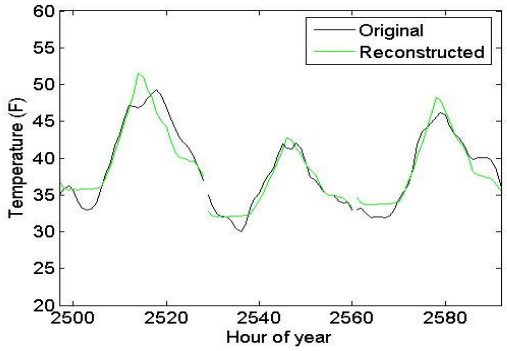
Self-organizing Map (SOM) is a powerful neural network method for the analysis and visualization of high dimensional data. It is also called Kohonen map because it was first introduced by Professor Teuvo Kohonen. This method is used for clustering and projecting of high dimensional data into a usually 1 or 2 dimensional map to reveal the data structure which is not explicit by visual inspection. By defining nodes in a map, SOM can be trained to cluster similar nodes together to represent neighborhood relationships between data items.



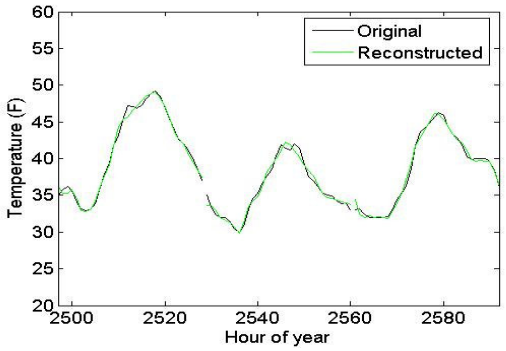
(a) Reconstruction with 1 coefficient



(b) Reconstruction with 2 coefficients



(c) Reconstruction with 3 coefficients

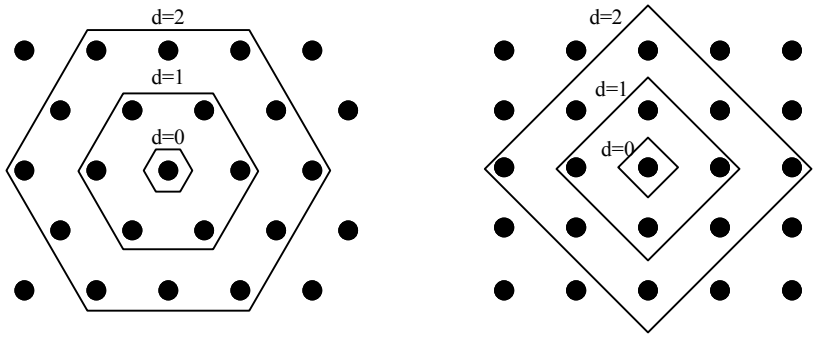


(d) Reconstruction with 10 coefficients

Figure 3.6. Comparison of original outdoor air temperature profiles with reconstructed temperature profiles using first 1, 2, 3 or 10 coefficients for three consecutive days.

Network Structure

A SOM consists of components called nodes or neurons. Each neuron has an associated weight vector of the same dimension as the input data vectors and a position in the map topology. The usual connection of neurons is in a hexagonal or rectangular grid as shown in Figure 3.7. The neurons can also be in a random pattern.



(a) Hexagonal grid (b) Rectangular grid

Figure 3.7. SOM neuron topology and the neighboring neurons to the centermost neuron for distance d=0, 1 and 2 (Vesanto, etc. 1999).

Neighborhood of each neuron contains all nodes within a distance d. The two diagrams show the neighboring neurons of the centermost neuron for distance d =0, 1 and 2.

Network Training

The goal of training in the SOM is to make different parts of the network responding similarly to certain input patterns. The weights of the neurons are initialized

to small random values. Usually the random values are within the variation of input data set for a faster training because the initial weights already give good approximation of SOM weights. The training utilizes competitive learning. When a training example is fed into the network, its distance, usually Euclidean distance, to every weight vector is computed. The neuron with weight vector most close to the input is called the best matching unit (BMU). For example, the distance between a neuron with weight $w\{w_1, w_2, ..., w_n\}$ and an input sample $v\{v_1, v_2, ..., v_n\}$ is given by:

$$dist = \sqrt{\sum_{i=1}^{n} (v_i - w_i)^2} \tag{3.6}$$

This gives a good measurement of how similar the two set of data are to each other. The weight of the BMU and its neighboring neurons in the same neighborhood within a distance d in the SOM grid are adjusted towards the input vector. The magnitude of the change varies with time and distance from the BMU. The update formula for a neuron with weight vector $w(t)$ is:

$$w(t+1) = w(t) + \Theta(d,t)L(t)[v(t) - w(t)] \tag{3.7}$$

where $L(t)$ is a decreasing learning rate. It can be any decay function like an exponential decay function. $\Theta(d,t)$, the neighborhood effect function, depends on the grid distance between the BMU and neuron w. The definition of $\Theta(d,t)$ is based on the principal that

neurons which are closer to BMU are influenced more than farther neurons. In the simplest form, it is one for all neurons close enough to BMU and zero for the others, but a Gaussian function is a common choice. The neighborhood effect function shrinks with time. At the beginning of iteration, the neighborhood is broad. The self-organizing takes place on the global scale. With the progress of the iterations, the neighborhood shrinks to just a couple of neurons. The weights are converging to local estimates until eventually the neighborhood is just the BMU itself and the BMU achieves a desired stable condition. This process is repeated for each input data sample.

Map Visualization

The U-matrix (unified distance matrix) (Ultsch and Siemon 1990) is an important visualization method. U-matrix gives the Euclidean distances between any two neighboring map neurons and the average Euclidean distance from a neuron to all its neighboring neurons at $d = 1$. U-matrix representation of the Self-organizing Map visualizes the distances between the neurons. The distance between the adjacent neurons is calculated and presented with different colorings between the adjacent neurons. A color with a large value between the neurons in U-matrix corresponds to a large distance and thus represents a boundary between clusters in the input space. A color with a small value indicates that a cluster (or neighborhood) with similar input data set exists. Small value color areas can be thought as clusters and large value color areas as cluster separators. This can be a helpful presentation when one tries to find clusters in the input data without having any priori information about the clusters.

Figure 3.8 is an example of U-matrix representation of SOM. The training samples are TMY2 (NREL, 1995) daily climate data in Newark, NJ. This sample data set contains 365 elements where each element is a 3-dimensioanal vector composed of daily average outdoor air temperature, daily average dew-point, and daily average solar radiation.

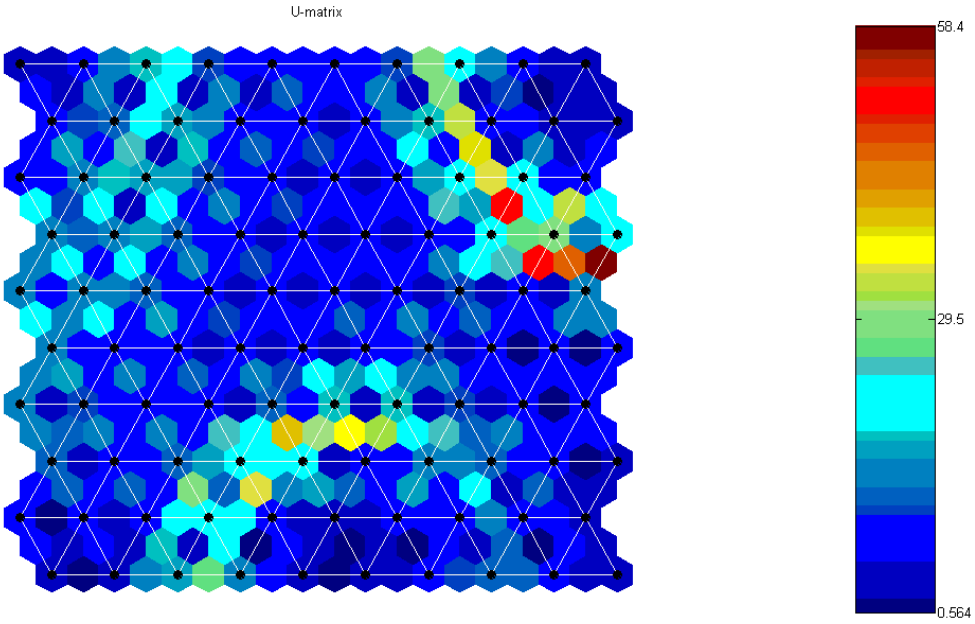


Figure 3.8. U-matrix representation of the Self-organizing Map.

In Figure 3.8, the neurons of the SOM are marked as black dots. The map has a 10 by 10 hexagonal grid topology. This representation reveals that there are a separate cluster in the upper right corner and a cluster in the lower central area of this representation. The clusters are separated by colors with high values. This result was achieved by unsupervised learning, that is, without human intervention. Training a SOM

and representing it with the U-matrix offer a fast way to get insight of the data distribution.

Neighborhood Classification

In the previous example, we defined a 10X10 grid map and trained the map with Newark TMY2 weather data. The U-matrix of SOM suggests 3 clusters (or neighborhoods) exist in the input sample data set which means the 365 days may be divided into 3 neighborhoods based on the daily weather similarities. To find the neighborhoods and days in each of them, it is necessary to define a 1X3 map where each of the three neurons represents a neighborhood. By training the map with the same weather data set. The associated weight vector of each neuron would then be updated to represent a cluster of the input data set respectively. 3 different neighborhoods can be determined by finding the BMU from the three neurons for each day.

Overview of Methodology

Determination of Significant Day Characteristics

A daily energy use model will be developed using wavelet coefficients as inputs to the neural network to determine the significant coefficients that affect building energy performance most. These coefficients are used as inputs to classify days into neighborhoods by SOM. Wavelet analysis of outdoor air temperature, dew point temperature, and solar radiation will be performed. Wavelet transform, instead of

traditional Fourier transform, has been chosen for feature extraction of the daily weather profiles because of the localization properties of wavelet analysis.

Hourly Energy Use Model

Once the daily building energy use regressors (significant wavelet coefficients) have been determined, the weights associated with the coefficients are determined as the derivatives of energy use with respect to the coefficients. Neighborhoods would then be classified based on the distances between the weighted daily regressor vectors. Prior to this, decorrelations between regressors must be considered to avoid interference effects.

The number of neighborhoods is determined by building constructions and climatic conditions. When an adequate number of neighborhoods are classified, an hourly building energy use prediction model will be developed for each of them.

The fully procedure to develop hourly neural network model is shown in Figure 3.9 by the following steps:

Step1: Hourly weather and building energy use data file generation

Step2: Apply discrete wavelet transform (DWT), develop daily energy use model and determine significant wavelet coefficients

Step3: Use SOM to classify neighborhoods

Step4: Develop hourly neural network model for each neighborhood

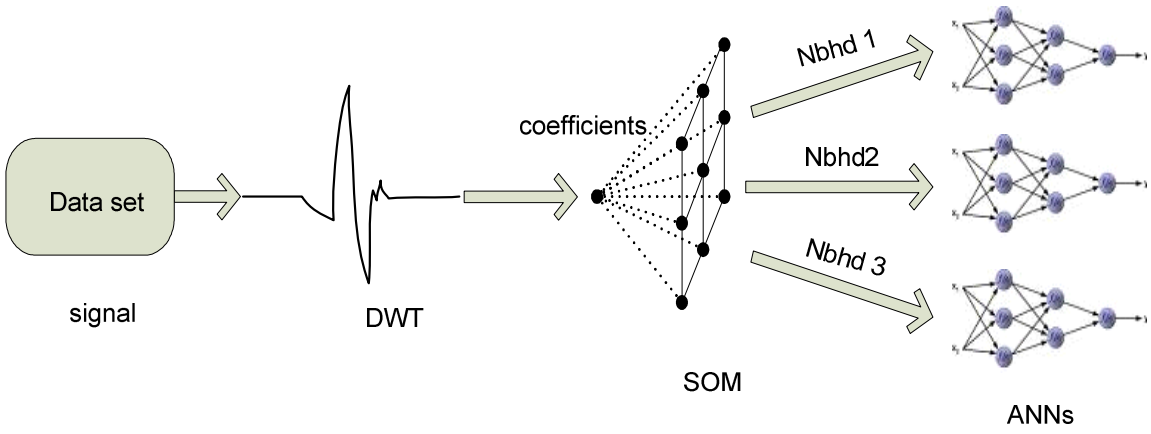


Figure 3.9. Hourly energy use prediction model.

Input variables of hourly neural network, such as weather variables, time stamps, and time-lagged variables are building dependent. The selection of input variables is a time consuming process, which requires consideration of both physical basis and operation schedule of the building.

Uncertainty Analysis

Traditionally, uncertainty of energy use prediction and energy saving estimates is determined by global estimates such as the overall RMSE. In this dissertation, a new approach based on "local" model behavior is presented. The "distance" concept is introduced to find the nearest neighboring days for a particular day by meteorological condition, and the energy prediction uncertainty for the particular day is determined by the distribution of modeling errors for these days. The energy use baseline model developed from pre-retrofit data can be used to estimate building energy savings

potential after installation of energy conservation measures. The daily energy saving for a post-retrofit day is the difference between measured post-retrofit energy use and baseline energy use. Building energy saving uncertainty will be estimated by finding the error distribution of the nearest neighboring days in the pre-retrofit period for the post-retrofit day. By this way, the uncertainty is determined by "local" prediction behaviors rather than global statistical indices.

Summary

In this chapter, the background knowledge of discrete wavelet transform (DWT) was introduced with a simple example. DWT is an important analysis tool which has been used for feature extraction. In this chapter, the SOM structure, training process, and map visualization were introduced also. An example of neighborhood classification of Newark TMY2 days was illustrated. The skeleton of methodology using discrete wavelet transform, Self-organizing Map, and neural network for building energy analysis has been overviewed. In the next chapter**,** development of daily energy use model using wavelet analysis will be presented.

CHAPTER IV

DEVELOPMENT OF DAILY ENERGY USE MODEL USING WAVELET ANALYSIS

Most statistical energy use models use algebraic daily average of weather components, such as outdoor air temperature, dew point temperature, and solar radiation as independent variables to simulate building daily energy uses. It is true if the building energy performance demonstrates a linear or approximately linear response to these variables. For large multi-zone buildings or buildings with complicated systems, variables other than the daily averages would be considered. As introduced in the previous chapters, wavelet coefficients are coefficients of wavelet transfer of input signal at different time locations and frequency levels. Wavelet coefficients at a certain level and location can explain variation of weather profiles well. Therefore, use the wavelet coefficients, combined with nonlinear model, would be a powerful tool to simulate building daily energy use. In this chapter, discrete wavelet transform is performed to find wavelet coefficients of the daily weather components. Daily energy use neural network models developed for an actual building case and four synthetic cases are studied to serve as examples for finding the most appropriate predictors from wavelet coefficients.

Description of Actual Building Case Study Data

Building and Data Introduction

The approach described above has been applied to data from a large campus building, Zachry Engineering Center, the one that was carefully selected for use in the energy prediction competition (Haberl and Thamilseran, 1996). Table 4.1 lists general information about the Zachry building. The building was constructed in the early 1970s. The primary retrofit to the building was to replace the existing constant volume air distribution systems with variable volume air distributions systems. The daily chilled water use (E_c) is the dependent variable of the model, while the independent variable set is comprised of the outdoor air temperature (T_{out}), dew point temperature (T_{dew}), and global horizontal solar radiation (I_{sol}). The data set contains the information of a total of 234 days.

Day Type Definitions and Data Preprocessing

The primary functions of Zachry building are for research and teaching activities. Therefore, building energy consumptions have heavy dependence on the operation schedule. Days are therefore divided into 3 different types based on the operation schedule:

- Day type I: weekends and holidays
- Day type II: days other than type I and type III
- Day type III: all week days during spring and fall semesters except holidays

Table 4.1. General Information about the Zachry Engineering Center, Texas A&M University [Thamilseran 1999]

TEXAS A & M UNIVERSITY: Zachry Engineering Center

Building Envelope:
- 324,400 sq.ft.
- 3-1/2 floors and a ground floor level, erected in 1973, classes, offices, labs, computer facility, and clean rooms for solid Electronics
- Walls: cement block
- Windows: 22% of total wall area
- single pane with built-in-place vertical blinds
- roof: flat

Building Schedule:
- classrooms and labs: 7:30 am to 6:30 pm weekdays
- offices: 7:30 am to 5:30 pm weekdays
- computer facility: 24 hrs/day

Building HVAC:
- 12 variable volume dual duct AHUs (12-40 hp)
- 3 constant volume multizone AHU (1-1hp, 1-7hp,1-10hp)
- 4 constant volume single zone AHU (4-3hp)
- 10 fan coils (10-0.5 hp)
- 2 constant volume chilled water pump (2.30hp)
- 2 constant hot water pump (2.20 hp)
- 7 misc. pumps (total of 5.8hp)
- 50 exhaust fans (50-0.5hp)

HVAC schedule:
- 24 hrs/day

Lighting: fluorescent

Retrofits Implemented:
- control modifications to the dual duct systems
- variable volume dual duct systems

Other Information:
- EMCS system to control HVAC was also installed along with the retrofits

Date of retrofits:
- date of completion for VAV and control modifications to the dual duct system: 3/30/91

With the definition of day type, building internal gain which exhibits strong diurnal pattern with respect to day type is not necessary to be treated as dependent variable for the model. Energy simulation model will then be developed for each day type.

In most large commercial buildings, a major portion of the latent load is from fresh air intake. Figure 4.1 is a scatter plot of cooling load E_c versus T_{out}. The scatter distribution of cooling energy use in cooling season shows that latent load is affected by humidity. Following the previous studies (Katipamula 1996; Katipamula et al. 1998), the term of $(T_{dew}-T_s)^+$, which is called effective dew point (ΔT_{dew}) in this dissertation, rather than T_{dew} in order to better capture humidity loads. T_s is the mean surface temperature of the cooling coil. The term is set to zero when it is negative and T_s is set to 55 °F for the buildings in this dissertation.

Selection of Wavelet and Decomposition Level

A time series signal input can be decomposed into approximation and detail coefficients at different levels. The selection of the appropriate wavelet filters and decomposition levels dependents on the specific problems. Some researchers would like to try all available filters and determine a most appropriate one. Some researchers do visual inspection of the signal characteristics and available wavelets, and select a wavelet that looks similar to dominant signal characteristics.

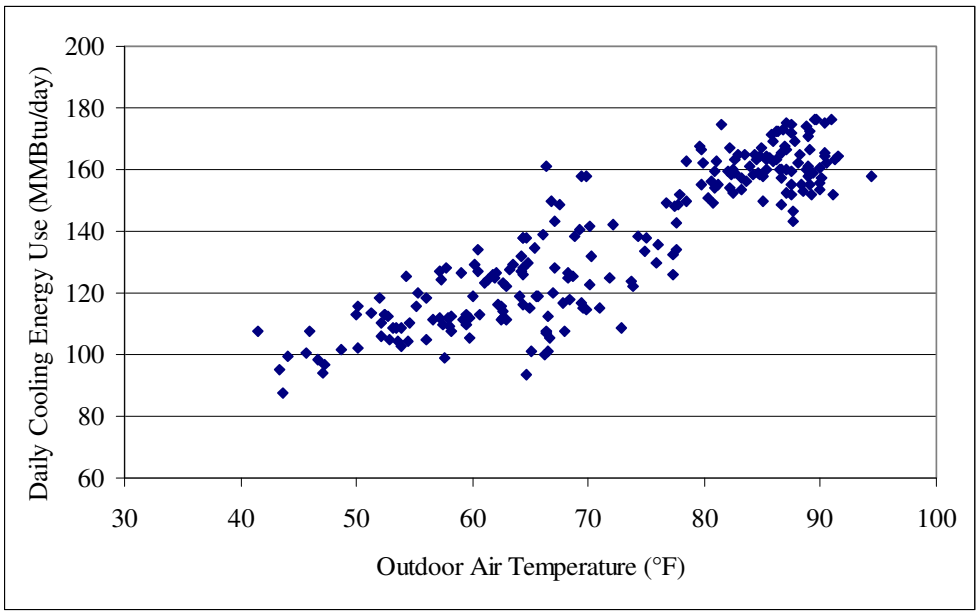


Figure 4.1. Scatter plot of thermal cooling load E_c versus T_{out} between 1/1/1990 and 11/27/1990 for Zachry Building.

Because of being no principled ways to select wavelets, we use Daubechies' wavelets for data analysis in current research. The input signals (or independent variables) for the daily energy model are T_{out}, ΔT_{dew} and I_{sol}. Because the variation of ΔT_{dew} is small and I_{sol} is symmetric during the day, the simplest wavelet Db1 is used for wavelet analysis of ΔT_{dew} and I_{sol}. For T_{out}, Db3 is selected because Db3 has longer filters than Db1; and the corresponding wavelet coefficients of T_{out} would account for the thermal storage effect caused by building mass.

Wavelet coefficients at different decomposition levels are also studied. Figure 4.2 displays average hourly T_{out}, ΔT_{dew} and I_{sol} for the year 1990 in College Station, TX. The T_{out} curve was created by separating daily outdoor air temperatures into 24 bins

corresponding to 24 hours of a day and then calculating mean value of each hour bin throughout the whole year. ΔT_{dew} and I_{sol} curves are determined in the same way.

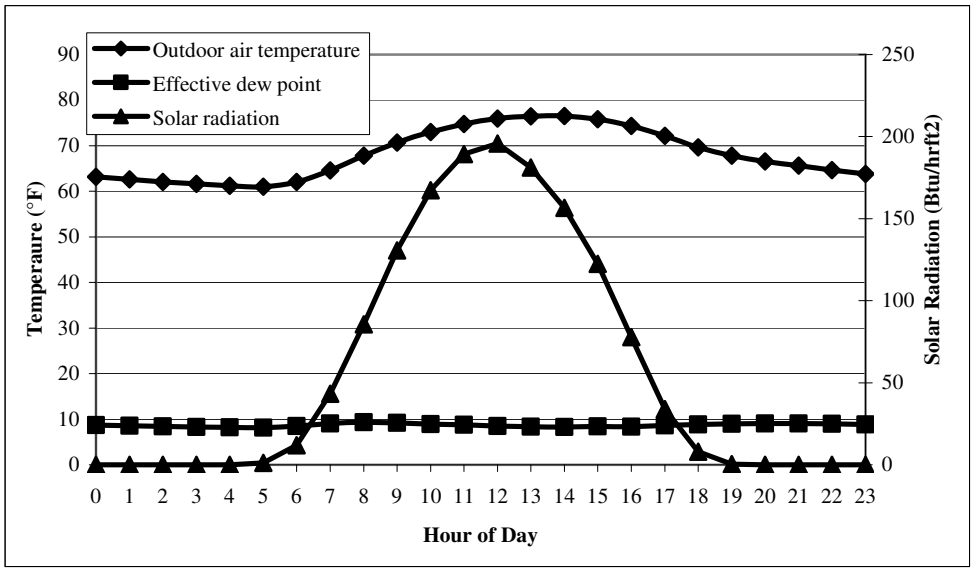


Figure 4.2. Average hourly weather data of a typical year, 1990, College Station, TX.

Figure 4.2 demonstrates a very stable dew point temperature curve, so use only daily average dew point would be enough to represent variation of the dew points during the day. Therefore, approximation coefficient at level $j = 5$ will be selected because the coefficient represents daily average dew point. The T_{out} curve shows an ascending trend in the first half of the day and descending trend in the second half of the day. The overall temperature fluctuation is small and close to the average daily temperature. Therefore, approximation coefficient a_5 at level $j = 5$ which represents average daily temperature and the detail coefficient d_5 at level $j = 5$ which represents the deviation to the average

will be studied. Solar radiation during the day obeys a well symmetric distribution from the Figure 4.2. There is no necessary to study wavelet coefficients in the morning and in the afternoon. The average I_{sol}, a_5 at level $j = 5$, is appropriate for daily energy simulation.

From the analysis of weather component curves, daily average $T_{out}, \Delta T_{dew}, I_{sol}$, detail coefficients of T_{out} at level $j = 5$ and their combinations will be used as predictors of neural network models, and the best predictors of the daily energy use model will be determined.

Training and Testing Data Sets

The whole dataset must be divided into a training set and a testing set in order to train the neural network model and test its performance. The building energy use data and climatic data of each day are put together in time sequence. Data at different seasons have different influences on building energy use. Random selection is then not a very good strategy. The training set must be most representative to cover the whole dataset. In this dissertation, instead of random selection, 2/3 of the days in the whole dataset are selected for training evenly in time sequence. The reset 1/3 are treated as testing dataset.

Parameters of Neural Network Model

The basic neural network structure is a feed forward network containing an input layer, a hidden layer, and an output layer. The output layer is building daily energy use. Generally, the number of neurons in hidden layer is roughly proportional to the dataset

size because more neurons are required to explain the inherent complicated variations in larger dataset. A neural network model with too many neurons may fit the training dataset very well but is not a generalized model. To find the optimal number of neurons, the coefficient of variation, $CV(RMSE)$, is employed. For example, a close training and testing $CV(RMSE)$ resulted from an neural network model with as few as possible neurons can be said an optimal choice. The coefficient of variation is defined as the following:

$$CV(RMSE)(\%) = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_{pred,i} - y_{data,i})^2}{n-p}}}{\overline{y}_{data}} \times 100$$

where, $y_{data,i}$ is measured energy use data for data point i; $y_{pred,i}$ is predicted energy use by the model for data point i; $\overline{y}_{data}$ is mean value of the measured energy use for the dataset; n is number of data point in the dataset and p is total number of regressor variables in the model.

For the purpose of determining the most appropriate input set of the neural network, several combinations of daily average T_{out}, ΔT_{dew}, I_{sol} and detail coefficients of T_{out} at level $j = 5$ are tested. The combinations are listed in Table 4.2. If the criterion for independent input selection is simply to best fit the data, the model with more inputs would be considered. However, the CV do not change much when extra inputs are

considered. The most practical way is to use inputs as few as possible but still well enough to account for energy performance variations. In this dissertation, up to three input variables are considered.

Table 4.2. Daily Energy Model Inputs

Case	Daily Energy Use Model Inputs
1	T_{out} at a_5
2	T_{out} at a_5, ΔT_{dew} at a_5
3	T_{out} at a_5, I_{sol} at a_5
4	T_{out} at a_5, ΔT_{dew} at a_5, I_{sol} at a_5
5	T_{out} at a_5, T_{out} at d_5
6	T_{out} at a_5, T_{out} at d_5, ΔT_{dew} at a_5
7	T_{out} at a_5, T_{out} at d_5, I_{sol} at a_5

* T_{out} at a_5 is referred to as approximation coefficient of T_{out} at level $j = 5$ and T_{out} at d_5 is referred to as detail coefficient of T_{out} at level $j = 5$ according to Figure 3.4.

Results of Actual Building Simulation

Based on the conditions defined previously, the neural networks are trained and tested. Table 4.3 has the coefficient of variation for each case that measures model performance.

Table 4.3. Coefficients of Variation of Daily Energy Model for Different Inputs

Case	CV_1 (%) training dataset	CV_2 (%) testing dataset	CV (%) whole dataset	Error (%) $(CV_1-CV_2)/CV_1$
1	7.264	6.570	7.042	-9.55
2	6.159	5.835	6.054	-5.26
3	6.833	6.686	6.785	-2.15
4	5.735	5.549	5.674	-3.24
5	6.639	7.098	6.794	6.91
6	6.057	6.044	6.052	-0.21
7	6.159	7.035	6.462	14.22

A method to validate whether the network is well developed or not is to compare CV between training set and testing set. The testing set and training set are independent to each other, so a close CV between training set and testing set means there is no over or under fitting of the model and the model is reliable to apply to more general cases. Table 4.3 indicates an error of 3.24% between two sets for case 4. Building daily E_c training and testing outputs are shown in Figure 4.3 and Figure 4.4 plotted against measured daily E_c. The Case 4, where the error between the two sets is small enough, has the smallest CV compared to other cases. It can be concluded that the average T_{out}, ΔT_{dew}, and I_{sol} are the best predictors for Zachry building daily cooling energy use.

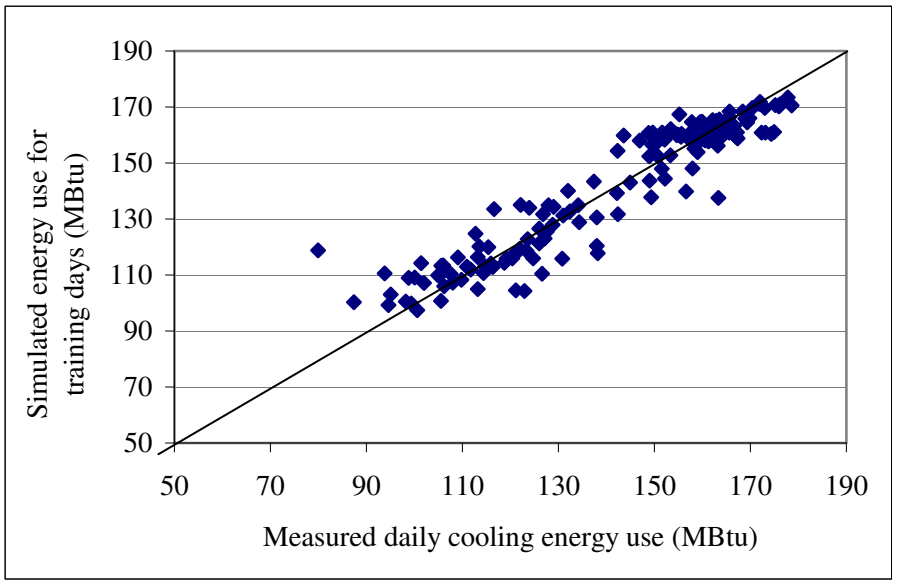


Figure 4.3. Measured and simulated E_c for training dataset of the Zachry Building.

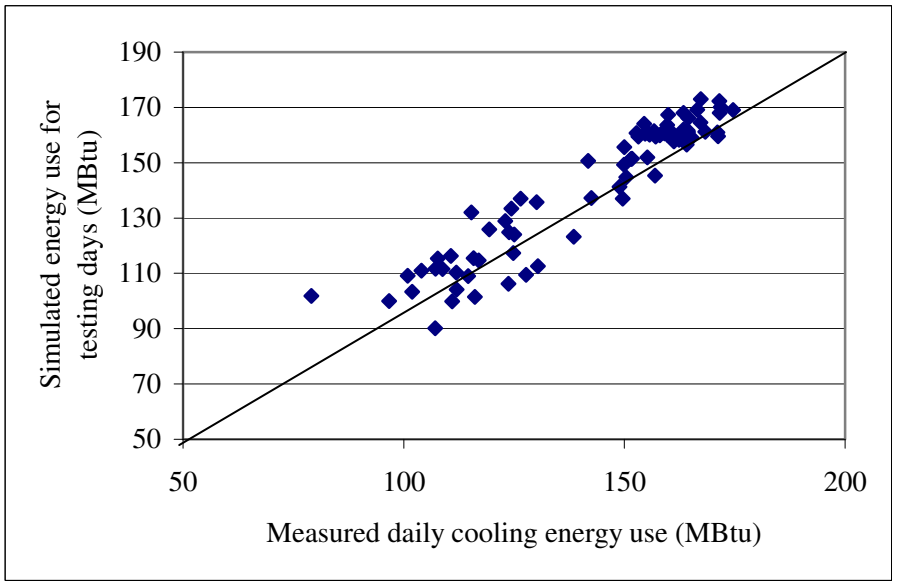


Figure 4.4. Measured and simulated E_c for testing dataset of the Zachry Building.

Synthetic Building Case Study

More buildings were studied in order to generally understand the relationship between wavelet coefficients of daily weather profile and building energy consumption. Four synthetic DOE2.1e building energy simulation models developed for four buildings in the framework of another study (Maor and Reddy 2008) are selected. The building envelope properties, systems, operating schedules and DOE2 DrawBDL pictures are listed in Appendix A.

Building Introductions

Large Hospital: The building is a seven story 315,000 ft^2, rectangular shaped building with 10 thermal zones. Floors 1 through 6 include 4 perimeter zones and one interior zone (all 6 floors assumed identical), and floor 7 also includes 4 perimeter zones and one interior zone. Building envelope properties, systems efficiencies, etc, are based on ASHRAE 90.1-2004 minimum requirements. Operating schedules (lighting, occupancy, etc.) are based mainly on data from ASHRAE 90.1-1989 and PG&E 2003 "Saving by Design Healthcare Modeling Procedures". Variable Air Volume (VAV) with hot water reheat was selected as secondary air system. ASHRAE 2003 "HVAC Design Manual for Hospital and Clinics" was used for additional design information.

Large Hotel: The building is a forty three story 619,200 ft^2, rectangular shaped building with 8 thermal zones. Floor 1 is lobby, shops and restaurants. Floor 2,3 and 4 accommodating conference rooms, banquet and offices. Floor 5 through 42 are guest rooms in a perimeter – core layout where the core includes corridors, shafts and service

rooms. Floor 43 accommodates mechanical rooms and service areas. Building envelope properties, systems efficiencies, etc. are based on typical design practices for the late 1980. Sections of ASHRAE 90.1-2004 minimum requirements used as well. Operation schedules (lighting, occupancy, etc.) are based mainly on data from ASHRAE 90.1-1989. Variable air volume (VAV) with hot water reheat air system was selected for the lobby, conference rooms, and other administrative areas. Four Pipe Fan Coils (FPFC) units (Chilled water and hot water) are used for guest rooms. The guest room floors core areas are served by a 100% OA, Dedicated Outdoor Air System (DOAS) which comprises a Reheat Fan System (RHFS).

Large Office: It is a large rectangular geometry, 588,000 ft^2, 17 story office building facility. The facility is a campus comprising of typical office and administration areas, and a mechanical penthouse to accommodate mechanical and electrical equipment. Building envelope properties, systems efficiencies, etc. are abased on typical design practice for the late 1990. Sections of ASHRAE 90.1- 2004 minimum requirements have been used. Operating schedules (lighting, occupancy, etc.) are based mainly on data from ASHRAE 90.1-1989. Variable air volume (VAV) with hot water reheat air system was selected for office and administration floors and single zone reheat was selected for mechanical room penthouse.

Large School: A large 229,700 ft^2 high school facility was designed to accommodate around 1500 students. Building envelope properties, systems efficiencies, etc, are abased on typical design practices for the late 1990. Sections of ASHRAE 90.1-2004 minimum requirements have been used as well. Operating schedules (lighting,

occupancy, etc.) are based mainly on data from ASHRAE 90.1-1989. A variety of secondary air systems were used for the design. These systems includes Four Pipe Fan Coils (FPFC) for classrooms, Variable air volume (VAV) with reheat for the common areas and administration, and Single zone reheat for auditorium, gymnasiums and cafeteria.

Day Type Definition

Large hospital and large hotel have a constant set point of zone cooling and heating thermostat throughout the year. Their occupancy schedule, lighting and equipment schedule do not change much from week days to weekend days and holidays. So, all the days are considered to be of the same day type, and one neural network model is used to simulate energy use for either of the buildings. Large office and large school have different operating schedules from hospital and hotel. Day types are defined as the following:

- Day type I: Sundays and holidays
- Day type II: Saturdays
- Day type III: Weekdays

Holidays are defined by DOE2.1e simulation program based on TMY2 data. A neural network model is developed for each of the three day types.

Climates Selection

Four climates associated with four metropolitan areas were selected: Houston, Denver, Newark, and San Francisco. Houston has the warmest and most humid climate conditions. Newark is cold in winter and humid in summer. San Francisco has the mildest and driest. Denver has coldest and dry climate. All these areas have the similar solar radiation through the year.

In the synthetic building simulations, TMY2 hourly weather data are used for all climates except Denver. No TMY2 data are available for Denver; data at Boulder Colorado is used instead as substitute for Denver metropolitan area.

Results of Synthetic Building Simulation

Neural network structure and parameters are the same as in the simulation of the Zachry building. Coefficients of variation for each case that measures model performance are listed in Tables 4.4-4.7. The best inputs are determined by two steps. First is to find the case with smallest *CV* for testing set. Neural network model that yields a smallest testing *CV* makes it the most accurate model because the testing set presenting a more general data set. Only model having the best fitting on testing set will be considered as the candidate. The second is to compare *CV* between testing set and training set. If *CV* for training set is smaller than or deviates not too much from the testing set, the model can be deemed as well-developed and can apply to other data. Table 4.8 lists the best predictors for all these cases.

Table 4.4. *CV* of Cooling Energy Use Modeling for a Large Hospital

Case #	*CV* (%)		
	Training set	**Testing set**	**Whole dataset**
Houston			
1	9.96	9.74	9.89
2	4.03	4.23	4.10
3	6.90	7.12	6.98
4	**3.66**	**4.24**	**3.86**
5	7.81	8.28	7.97
6	3.26	4.00	3.53
7	6.32	6.29	6.31
Denver			
1	18.89	17.51	18.45
2	11.30	10.06	10.91
3	16.96	17.37	17.10
4	11.02	9.55	10.57
5	18.28	16.70	17.78
6	**10.07**	**11.30**	**10.49**
7	13.71	14.20	13.87
Newark			
1	22.34	22.94	22.54
2	**6.09**	**6.27**	**6.15**
3	15.57	16.92	16.02
4	6.14	7.33	6.55
5	17.04	18.21	17.43
6	7.90	8.72	8.17
7	14.92	17.36	15.76
San Francisco			
1	10.26	12.86	11.17
2	8.55	9.72	8.95
3	9.69	50.07	29.64
4	8.34	9.82	8.85
5	8.37	8.69	8.47
6	**5.75**	**6.28**	**5.93**
7	8.04	8.72	8.26

Table 4.5. *CV* of Cooling Energy Use Modeling for a Large Hotel

Case #	*CV* (%)		
	Training set	**Testing set**	**Whole dataset**
Houston			
1	12.37	12.99	12.58
2	11.57	12.23	11.80
3	11.81	12.64	12.1
4	**10.98**	**10.73**	**10.9**
5	10.78	11.84	11.15
6	10.11	11.71	10.68
7	11.73	13.13	12.22
Denver			
1	23.58	25.15	24.11
2	19.95	22.45	20.81
3	21.26	23.29	21.96
4	**19.32**	**21.61**	**20.11**
5	21.17	22.93	21.77
6	18.83	21.74	19.84
7	18.88	22.69	20.23
Newark			
1	22.57	22.31	22.48
2	17.32	17.36	17.33
3	20.49	21.52	20.83
4	16.38	18.11	16.97
5	18.17	19.6	18.65
6	**15.66**	**17.17**	**16.17**
7	17.87	18.89	18.21
San Francisco			
1	15	16.35	15.47
2	14.79	16.15	15.26
3	14.04	15.77	14.65
4	14.31	15.64	14.77
5	13.71	15.37	14.29
6	**14.21**	**15.21**	**14.55**
7	13.38	17.25	14.8

Table 4.6. *CV* of Cooling Energy Use Modeling for a Large Office

Case #	*CV* (%)		
	Training set	**Testing set**	**Whole dataset**
Houston			
1	11.29	12.01	11.55
2	9.18	30.54	19.44
3	12	13.17	12.42
4	11.21	20.21	14.95
5	10.87	15.74	12.77
6	**6.18**	**11.17**	**8.26**
7	9.12	18.51	13.15
Denver			
1	19.74	17.13	18.85
2	19.35	16.32	18.33
3	18.91	19.28	19.05
4	20.72	20.34	20.6
5	**17.71**	**16**	**17.12**
6	21.43	21.15	21.34
7	15.66	42.31	28.26
Newark			
1	16.31	19	17.2
2	27.25	28.53	27.66
3	15.32	18.33	16.32
4	27.62	28.93	28.05
5	15.63	18.78	16.68
6	18.77	19.68	19.06
7	**15**	**18.06**	**16.02**
San Francisco			
1	14.81	12.6	14.09
2	14.53	13.13	14.07
3	14.16	12.86	13.73
4	13.82	14.13	13.93
5	8.98	9.33	9.1
6	**8.44**	**7.29**	**8.06**
7	8.25	11.2	9.37

Table 4.7. *CV* of Cooling Energy Use Modeling for a Large School

Case #	*CV* (%)		
	Training set	**Testing set**	**Whole dataset**
Houston			
1	35.77	36.78	36.14
2	34.22	37.86	35.56
3	32.06	37.22	33.99
4	**30.18**	**34.23**	**31.68**
5	34.82	37.11	35.66
6	32.66	38.18	34.72
7	34.58	116.77	74.87
Denver			
1	22.92	19.57	21.86
2	18.62	16.97	18.09
3	22.08	21.08	21.75
4	**18.18**	**16.02**	**17.49**
5	21.99	19.1	21.07
6	18.47	16.78	17.92
7	21.34	19.87	20.86
Newark			
1	24.35	22.23	23.73
2	14.54	14.81	14.63
3	21.5	20.97	21.35
4	**13.09**	**15.38**	**13.84**
5	21.99	20.05	21.42
6	14.66	16.53	15.26
7	20.26	19.57	20.06
San Francisco			
1	35.14	36.82	35.75
2	35.08	39.41	36.68
3	27.15	32.94	29.33
4	26.95	33.69	29.51
5	32.99	33.67	33.24
6	32.84	35.13	33.68
7	**26.69**	**31.57**	**28.51**

Table 4.8. Best Daily Cooling Energy Use Predictor for Synthetic Buildings

Large hospital	**Houston**	T_{out} at a_5, ΔT_{dew} at a_5, I_{sol} at a_5
	Denver	T_{out} at a_5, T_{out} at b_5, ΔT_{dew} at a_5
	Newark	T_{out} at a_5, ΔT_{dew} at a_5
	San Francisco	T_{out} at a_5, T_{out} at b_5, ΔT_{dew} at a_5
Large hotel	**Houston**	T_{out} at a_5, ΔT_{dew} at a_5, I_{sol} at a_5
	Denver	T_{out} at a_5, ΔT_{dew} at a_5, I_{sol} at a_5
	Newark	T_{out} at a_5, T_{out} at b_5, ΔT_{dew} at a_5
	San Francisco	T_{out} at a_5, T_{out} at b_5, ΔT_{dew} at a_5
Large office	**Houston**	T_{out} at a_5, T_{out} at b_5 ΔT_{dew} at a_5
	Denver	T_{out} at a_5, T_{out} at b_5,
	Newark	T_{out} at a_5, T_{out} at b_5, I_{sol} at a_5
	San Francisco	T_{out} at a_5, T_{out} at b_5, ΔT_{dew} at a_5
Large school	**Houston**	T_{out} at a_5, ΔT_{dew} at a_5, I_{sol} at a_5
	Denver	T_{out} at a_5, ΔT_{dew} at a_5, I_{sol} at a_5
	Newark	T_{out} at a_5, ΔT_{dew} at a_5, I_{sol} at a_5
	San Francisco	T_{out} at a_5, T_{out} at b_5, I_{sol} at a_5

Building energy performance is influenced by not only climate, but also building construction and operating schedule. No generalized rule for the determination of significant wavelet coefficients can be derived. Case by case study is necessary for specific building energy modeling.

Summary

Discrete wavelet analysis of daily weather component profiles and the selection of wavelet and wavelet decomposition level have been introduced in this chapter. The

daily energy use model used to determine significant wavelet coefficients are developed. The significant wavelet coefficients are different for different buildings and climates. In the next chapter, using significant wavelet coefficients to classify neighborhood by Self-organizing Map will be presented.

CHAPTER V

NEIGHBORHOOD CLASSIFICATON

Traditionally, building energy use statistical models such as simple regression model, multivariate linear (MLR) model and neural network model (ANN) etc. are used to capture the global behavior of the dependent variable using the entire data set where the entire range of independent variables encompassing disparate conditions. A global model needs to account for very different consumption patterns throughout the analysis period. The simplicity of the application of these global models would compromise simulation accuracy, or in the other word, introduce higher prediction uncertainty comparing to the models developed on a relatively "local" range identified by the similar system behaviors. In this dissertation, the idea that to classify the days, in which building exhibits similar system behaviors, into the same neighborhood is proposed. The energy use baseline would then be developed for each neighborhood in order to achieve a higher applicability to energy performance for the days within a neighborhood. This approach which is based on the local system performance is more realistic, and hence more robust and credible, than the global models currently used.

Regression Variable De-correlations and Weights

In the previous chapter, the neural network daily energy use model was used to determine the most significant wavelet coefficients of climate components for daily cooling energy use. But the influences of these significant coefficients, called weights in

this dissertation, on energy use are still unknown. The weights contain physical meaning in the energy use model and are critical in finding similarities between the days when wavelet coefficients are used to characterize the meteorological behavior of a day.

Weights Calculation Method

The simplest way to get the weights is to use a multivariate linear model where wavelet coefficients are regressors and daily energy use is output. The resulting coefficients of MLR are weights of corresponding wavelet coefficients. Another method is to find the derivative of building energy use with respect to each regressor in the daily neural network model. The derivatives are weights that we are seeking. This process is more complicated but keeps the integrity of the weights and significant coefficients in the same model. The former method is used for weights determination is this chapter. The latter method will be applied in Chapter VIII.

De-Correlation of Collinearity

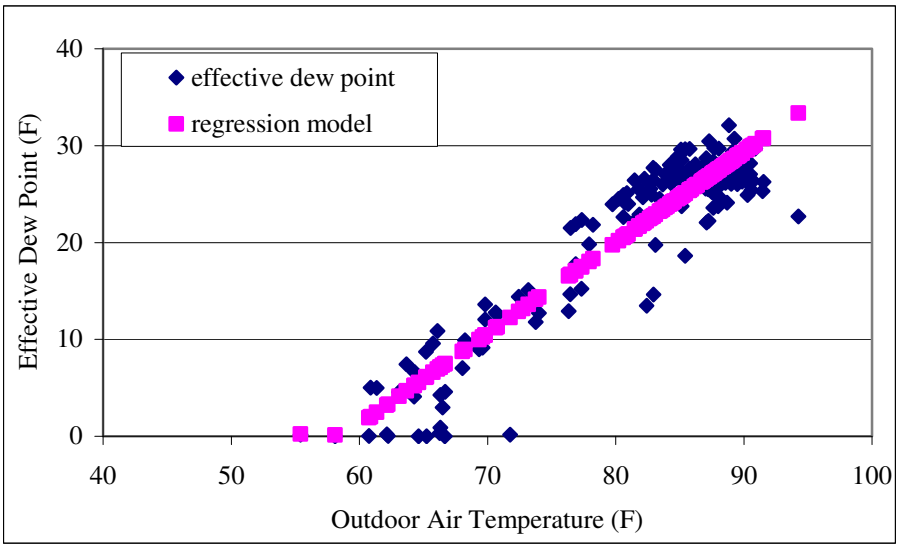
A critical issue with multivariate models in general is the collinear behavior between regressors. In building energy use modeling, colinearity lies in the significant correlation between outdoor air temperature and dew point temperature. If we ignore this, the energy use model may have physically unreasonable internal parameter values, but continue to give reasonable predictions. However, the derivatives of the response variable with respect to the regressor variables (or weights of regressor variables) can be very misleading. One variable may “steal” the dependence from another variable, which

will affect weights assigned to the different regressors. If this correlation between regressors is linear, a standard way is to use principle components analysis (PCA) to minimize the advert affects. Unfortunately, building energy use modeling that shows a nonlinear correlation between regressors benefits little from PCA. Figure 5.1 illustrates the nonlinear correlation between T_{out} and ΔT_{dew}.

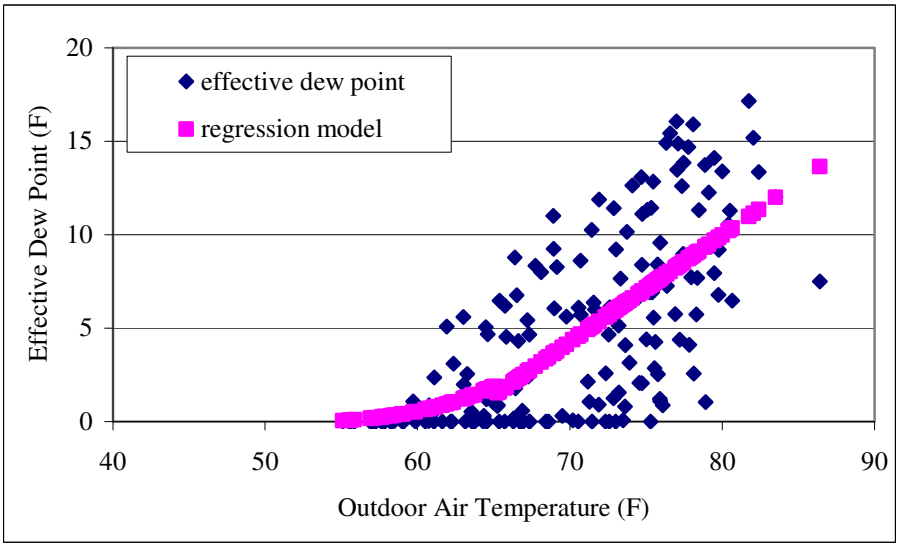
To mitigate the correlation, we proposed fitting a model to ΔT_{dew} vs T_{out}, and using residuals of this model ($\mathrm{Res}T_{dew}$) instead of ΔT_{dew} as the regressor for energy use modeling. Thus:

$$\mathrm{Res}T_{dew} = (T_{dew} - 55)^{+}_{meas} - (T_{dew} - 55)^{+}_{model} = \Delta T_{dew,meas} - \Delta T_{dew,model}$$

where $\Delta T_{dew,model}$ may be either a simple regression model or an ANN model between $(T_{dew} - 55)^{+}$ and T_{out}. Examples of $\Delta T_{dew,model}$ are illustrated in Figure 5.1 and $\mathrm{Res}T_{dew}$ are plotted in Figure 5.2.

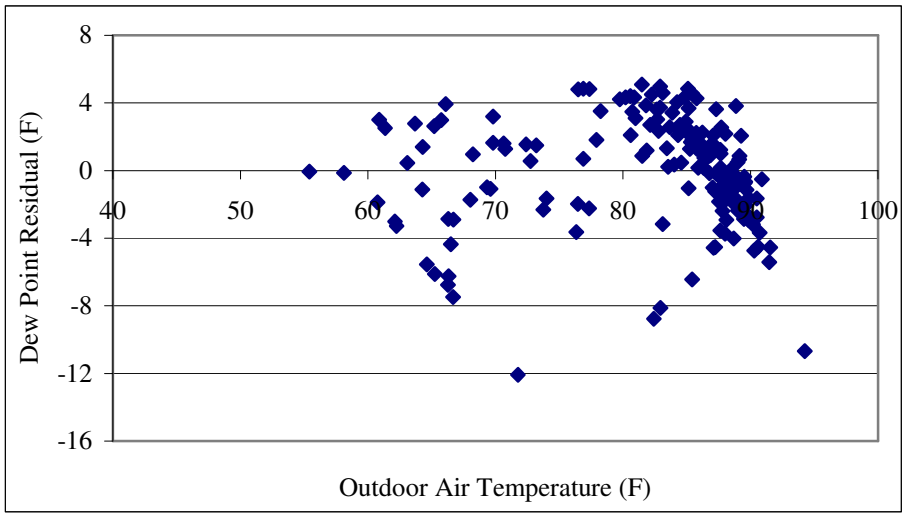


(a) Zachry building in College Station, TX.

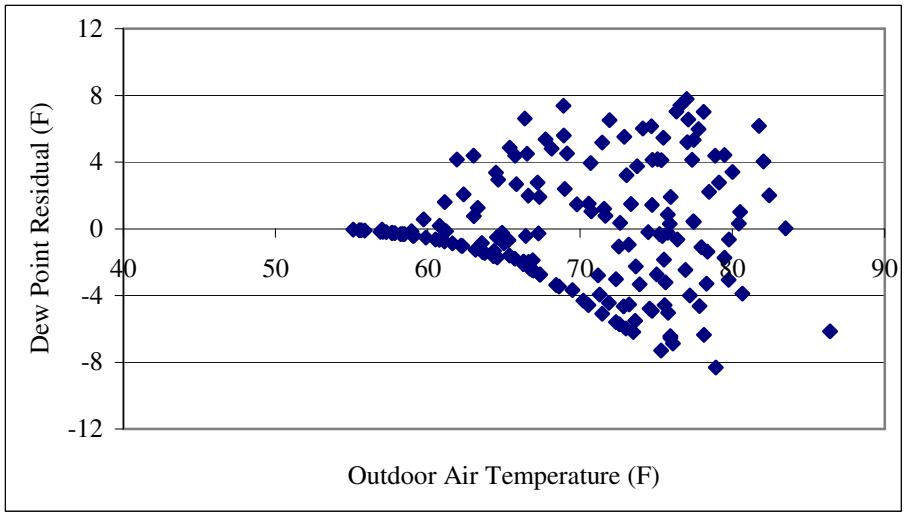


(b) Large hotel in Newark, NJ.

Figure 5.1. Correlation and regression model between T_{out} and ΔT_{dew}.



(a) Zachry building in College Station, TX.



(b) Large hotel in Newark, NJ

Figure 5.2. Plot of $\mathrm{Res}T_{dew}$ against T_{out}.

Weights Calculations

A sample weights calculation of the Zachry building is conducted. In daily energy use chapter, the significant wavelet coefficients for the Zachry building have been determined as T_{out} at a_5, I_{sol} at a_5 and ΔT_{dew} at a_5. After applying polynomial fitting of T_{out} at a_5 versus ΔT_{dew} at a_5, significant wavelet coefficients become T_{out} at a_5, I_{sol} at a_5 and ResT_{dew} at a_5. A multivariate linear modeling yields the following weights of the individual significant wavelet coefficients in Table 5.1.

Table 5.1. Weights of Significant Wavelet Coefficients for the Zachry Building Cooling Energy Use

Wavelet coefficients	T_{out} at a_5	ResT_{dew} at a_5	I_{sol} at a_5
Weights	1.3677	2.5092	0.0682

Although the sensitivity of the MLR model to T_{out} is not the biggest, its numerical value is high, and so its absolute impact on E_c is still the most significant. As to ResT_{dew} and I_{sol}, both have a roughly equal impact on daily E_c consumption by considering their numerical values.

Another example is a large hotel in Newark. Significant wavelet coefficients are T_{out} at a_5, T_{out} at d_5 and I_{sol} at a_5 determined by daily energy use model in the previous chapter. By applying a multivariate linear regression, the weights are as follows in Table 5.2.

Table 5.2. Weights of Significant Wavelet Coefficients for Large Hotel Cooling Energy Use

Wavelet coefficients	T_{out} at a_5	T_{out} at d_5	ResT_{dew} at a_5
Weights	0.5520	0.1158	3.8839

Neighborhood Classification for Meteorological Days

By feeding the significant wavelet coefficients and their weights of weather data for each day during a period into SOM, the days in this period would then be automatically classified into a number of neighborhoods. The days in the same neighborhood have similar meteorological characteristics and should have similar influence on building energy performance if they have the same day type. The Zachry building in College Station and a large hotel in Newark are used as examples to demonstrate how SOM classifies neighborhoods.

Neighborhood Classification for the Zachry Building in College Station

The Zachry building data set used in chapter IV contains 234 days of complete weather data. The U-matrix plot suggests 3-4 neighborhoods to be classified. 4 neighborhoods is a moderate number that represents weather variation of seasonal change as well as keeping enough days in each neighborhood for the energy use modeling. Table 5.3 lists the number of days in each neighborhood from SOM.

Table 5.3. Number of Days classified in Each Neighborhood for the Zachry Building

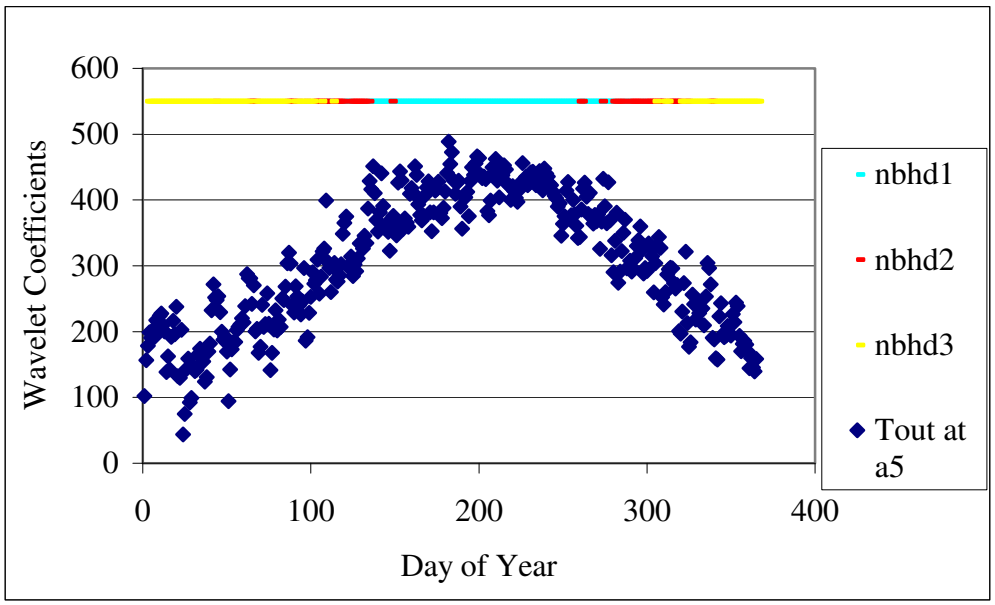
Wavelet coefficient	Neighborhood 1	Neighborhood 2	Neighborhood 3	Neighborhood 4	Total
T_{out} at a_5, I_{sol} at, a_5, ResT_{dew} at a_5	35 days	49 days	50 days	100 days	234 days

Neighborhood Classification for Large Hotel in Newark

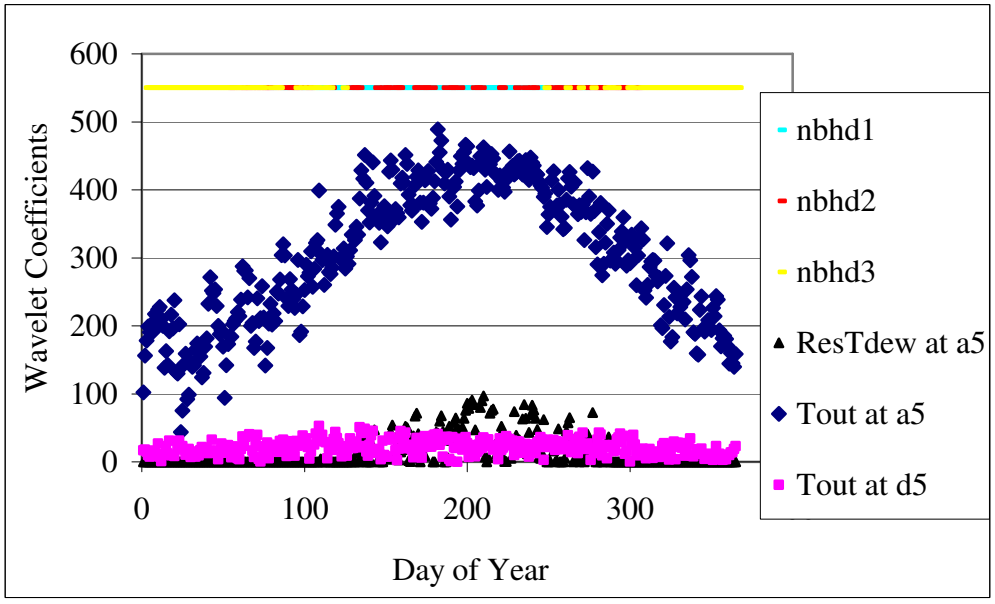
Using the same method, days are classified into 3 neighborhoods for the large hotel in Newark. Table 5.4 lists number of days in each neighborhood from SOM. Figure 5.3 illustrates the days in each neighborhood. Figure 5.3(a) shows the neighborhoods classified based on T_{out} at a_5 which means the average outdoor air temperature. Figure 5.3(b) shows the neighborhoods classified based on significant wavelet coefficients. The horizontal color bar in the figure represents the distinct neighborhoods classified for the 365 days.

Table 5.4. Number of Days Classified in Each Neighborhood for Large Hotel in Newark

Wavelet coefficient	Neighborhood 1	Neighborhood 2	Neighborhood 3	Total
T_{out} at a_5, T_{out} at d_5, ResT_{dew} at a_5	88 days	128 days	149 days	365 days



(a) Neighborhoods determined by T_{out} at a_5.



(b) Neighborhoods determined by T_{out} at a_5, T_{out} at d_5 and ResT_{dew} at a_5.

Figure 5.3. Neighborhood classifications for large hotel in Newark.

Summary

Dew point temperatures are found highly correlated to outdoor air temperatures. Regression model between the two variables used to eliminate the correlations and prevent weight "stealing" is necessary. Weights of the significant wavelet coefficient are calculated. Neighborhoods are determined by Self-organizing Map based on similarities of the days' characteristics defined by significant wavelet coefficients of the days and their weights. In the next chapter, the hourly energy use ANN model that was developed for each neighborhood will be discussed.

CHAPTER VI

BUILDING HOURLY ENERGY USE NEURAL NETWORK MODEL

In Chapter V, the method of neighborhood classification has been introduced to divide the data set into several neighborhoods based on the daily weather condition. Hourly energy use model developed in each of the neighborhood will be discussed in this chapter. The comparison between the energy use modeling based on the neighborhood classification and modeling without neighborhood classification will be performed.

Zachry Building Hourly Energy Use Modeling

Once the neighborhoods are classified, a nonparametric or parametric model can be developed in each of the neighborhoods. In this research, we adopt ANN models which have several redeeming qualities such as being able to handle spatial non-linearities in the response variable in a logical and automated manner.

Training Data Set for ANN Model

The training data set is the same as the data set used for daily energy use modeling in Chapter IV. According to the variation of weather data and their influence on energy use of the Zachry building, the weather data were divided into 4 neighborhoods which contain 34, 49, 50 and 100 days respectively. The first day is excluded from any neighborhood because time-lagged values are predictors of the model.

We still use 2/3 of the total data in each neighborhood for ANN training, and the rest are for model testing. The statistic index of *CV* is employed as the measurement of accuracy for the model. A close *CV* for training dataset energy modeling and testing dataset energy modeling indicates a well developed model with respect to the whole given data set.

ANN Parameters and Input Variables

The hidden layer of the feed-forward neural network has three neurons. The input layer is a 7-dimensional vector which contains 7 components: current hour outdoor air temperature T_{out}, effective dew point ΔT_{dew}, solar radiation I_{rad}, hour of the day, day type, previous hour outdoor air temperature $T_{out}^{'}$ and effective dew point $\Delta T_{dew}^{'}$. Previous hour $T_{out}^{'}$ and $\Delta T_{dew}^{'}$ are included in the input vector because of thermal mass effect of building envelope on systems.

Results and Comparison

Figure 6.1~6.4 have the simulated energy use data against measured energy use data for the Zachry Building with and without neighborhood classification. Table 6.1 compares the *CVs*. With neighborhood classification, the *CV* of 7.52% for training dataset energy modeling and *CV* of 8.2% for testing dataset energy modeling are both smaller than the corresponding *CV* of energy modeling without neighborhood classification in building energy analysis.

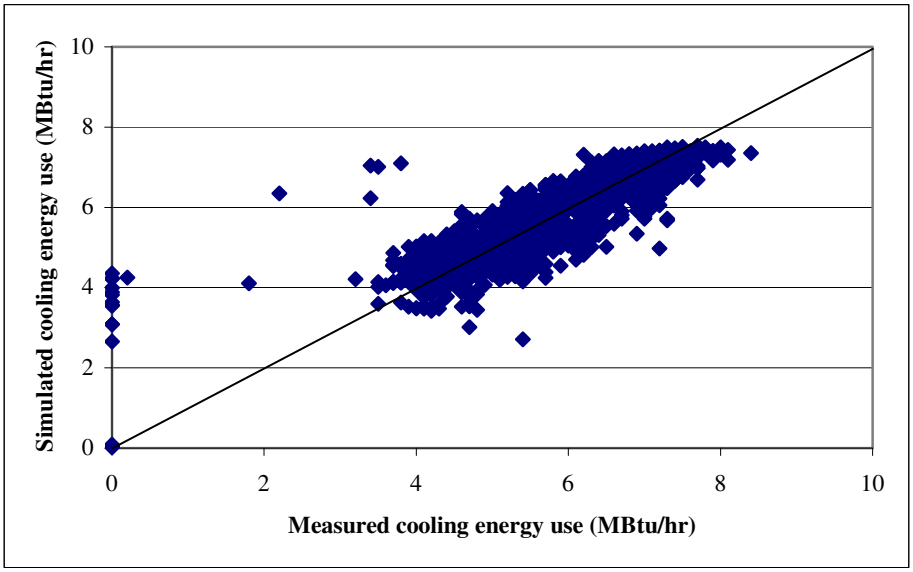


Figure 6.1. Simulated cooling energy use for training dataset with neighborhood classification for the Zachry Building.

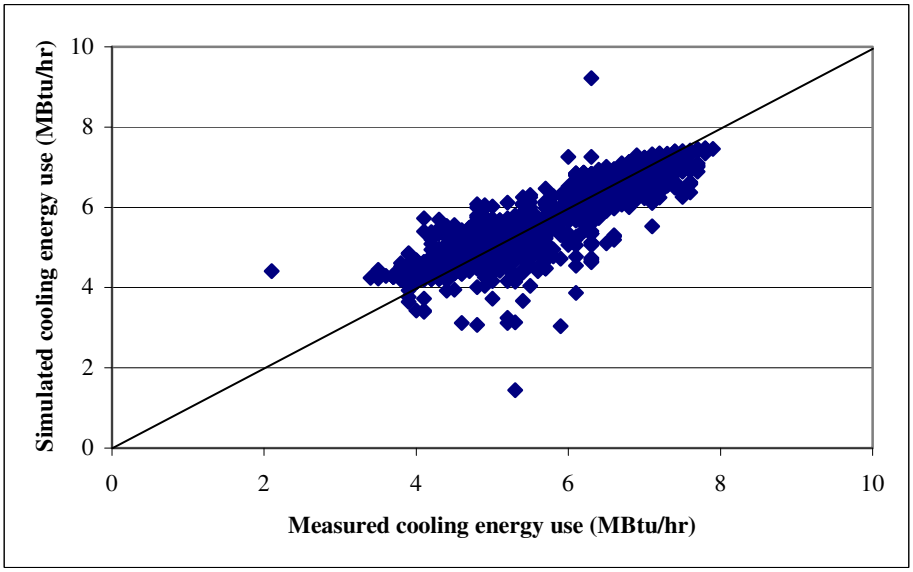


Figure 6.2. Simulated cooling energy use for testing dataset with neighborhood classification for the Zachry Building.

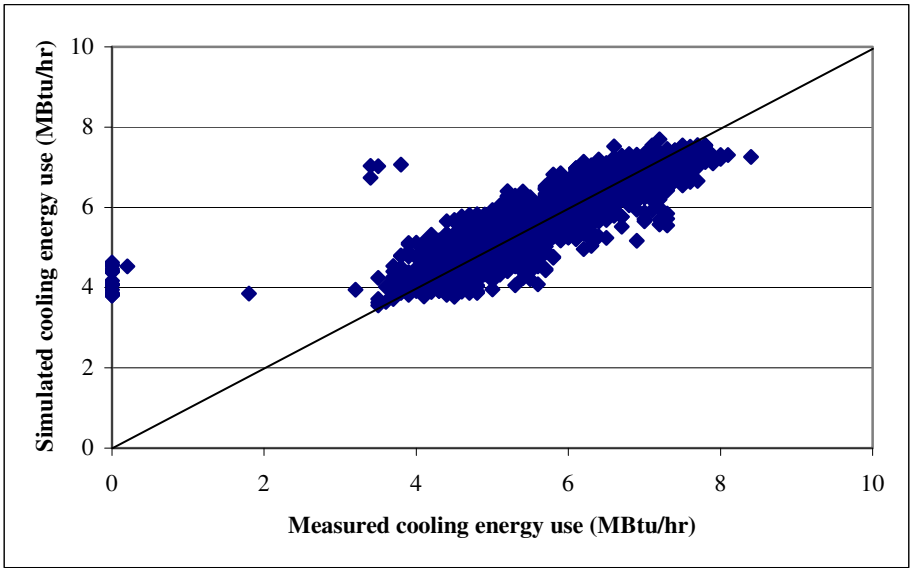


Figure 6.3. Simulated cooling energy use for training dataset without neighborhood classification for the Zachry Building.

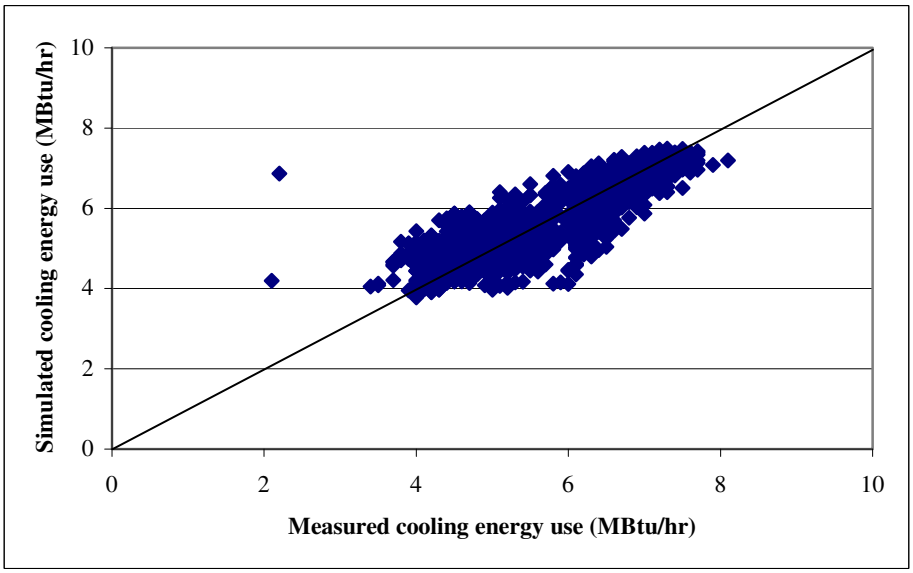


Figure 6.4. Simulated cooling energy use for testing dataset without neighborhood classification for the Zachry Building.

Table 6.1. Comparison of Modeling With and Without Neighborhood Classification

	With neighborhood classification					Without neighborhood classification
	Nbhd 1	Nbhd 2	Nbhd 3	Nbhd 4	Combined	
Training set *CV* (%)	5.20	10.43	11.92	4.94	7.52	8.20
Testing set *CV* (%)	5.45	10.52	12.30	4.54	7.50	8.32

Summary

Neighborhood-based hourly energy use ANN models are developed for energy use prediction in this chapter. The results between energy use modeling with and without neighborhood classification have also been compared. This simple comparison suggests that a baseline model developed by local system behavior is more reliable than a global energy model. In the next chapter, more comparisons will be performed for better understanding the proposed methodology.

CHAPTER VII

COMPARISON AND ANALYSIS

The previous chapters discussed the theory of the neighborhood based neural network model and its application. This chapter presents the application of the neighborhood based neural network model to dataset C from the Great Energy Predictor Shootout II (Haberl and Thamilseran 1996) and the comparison of this model to the Great Energy Predictor Shootout II winning entries for hourly energy use modeling and to Change-point model for daily energy modeling.

The Great Energy Predictor Shootout II

In order to evaluate many of the analytical methods in building energy use data analysis, an open competition was held by ASHRAE in 1993. The objective was to identify and compare the most accurate methods for building hourly energy use predictions based on limited amount of measured data (Kreider and Haberl 1994). Because of overwhelming response to this competition, a second Shootout was developed to compare how well different empirical models predict building energy and compare how those models can be used to calculate energy conservation retrofit savings in 1994.

Data Description

Two datasets were provided from two different buildings selected from about 100 monitored buildings that are part of the Texas LoanSTAR program. The first dataset which contains two files, C.trn and C.tst, was selected from energy consumption data for the Zachry Engineering Center located at Texas A&M University, College Station, Texas. The detailed information about this building has been introduced in Chapter IV. The second dataset which contains two files, D.trn and D.tst, is from the Business Building located at the University of Texas at Arlington, Texas. Data file C.trn is building pre-retrofit training dataset between 1/1/1990 0:00 and 11/27/1990 23:00. Data file C.tst is building post-retrofit testing dataset between 11/28/1990 0:00 and 12/31/1992 23:00. These two datasets contain independent variables (i.e., weather data and calendar time stamp) and the corresponding dependent variables (e.g., whole-building energy use). Portion of the dependent variables were removed from the training file. The independent variables that corresponded to the removed dependent data in the training file were used by the contestants to predict energy use for the removed periods. The predictions of energy use for the removal period in the training data set were then compared to the actual data to test the accuracy of the contestant's model. The provided independent weather variables are hourly outdoor air temperature T_{out}, relative humidity RH, solar radiation I_{sol} and wind speed V_{wind}. But some of the independent variables that corresponded to the removed dependent data were missing and required for filling by contestants.

Post-retrofit data file C.trn contains both independent and dependent variables from the post-retrofit period for the Zachry Engineering Center. The contestants were required to use their baseline models developed from training dataset to predict hourly baseline energy use and energy savings from the retrofitting. For a detailed data description, please refer to Thamilseran's dissertation (Thamilseran 1999).

In this dissertation, only whole building chilled water energy use E_c and hot water energy use E_h from the Zachry Engineering Center were studied and analyzed. Figure 7.1 and Figure 7.2 illustrate chilled water use and hot water use between 11/28/1990 0:00 and 12/31/1992 23:00. The predictions of the removed data shown as those discontinuities in the figures were used to evaluate the performance of energy use models.

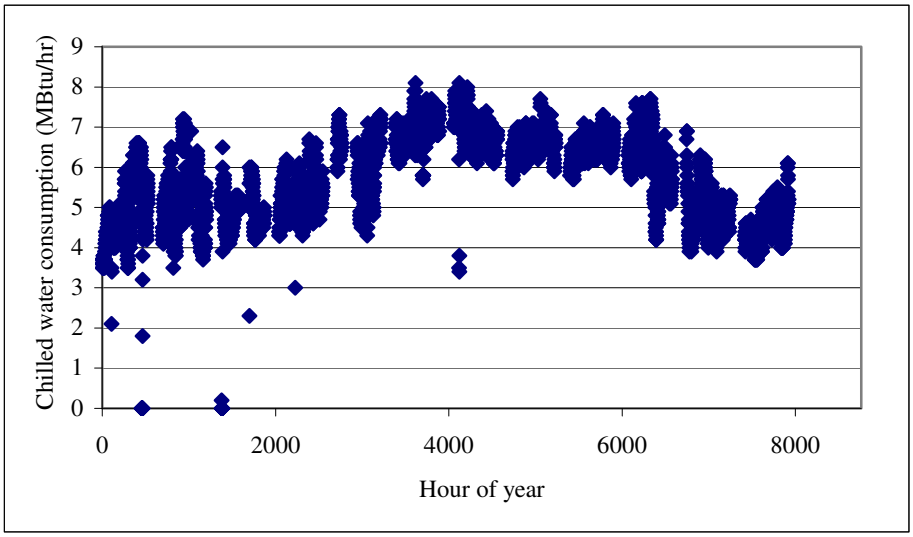


Figure 7.1. Cooling energy use data for Shootout II.

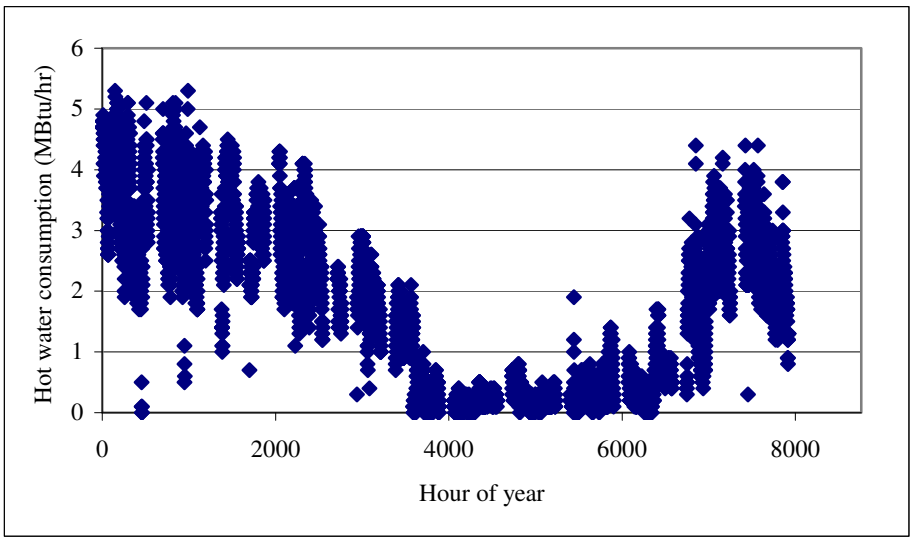


Figure 7.2. Heating energy use data for Shootout II.

Pre-processing of Shootout II Data

Data Inspection

There are totally 331 days between 1/1/1990 and 11/27/1990. For cooling energy use, only 234 days were provided with complete hourly chilled water use data. The removed 1864 hourly chilled water uses in the other 97 days were required to be predicted by the contestants. For heating energy use, only 236 days were provided with complete hourly hot water use data. The removed 1871 hourly hot water uses in the other 95 days were required to be predicted by the contestants. By visual inspection of the building energy use data, some outliers are excluded from dataset which are obviously not in the energy consumption trend. These outliers may be caused by metering errors or outage of the heating/cooling supply. Finally, 230 days containing complete cooling energy use data will be utilized to develop and test cooling hourly

energy use model and 231 days containing complete heating energy use data will be utilized to develop and test heating energy use model.

Data Filling of Missing Independent Variables

In this competition, different data filling methods were used by different contestants. For example, a method called “mean imputation” was employed by Dodier (Dodier and Henze 1996), in which the average value of a variable is substituted when the variable is missing. Jang adopted an autoassociative neural network as a preprocessor to replace the missing data (Jang et al. 1996).

Baltazar and Claridge (2002) studied the restoration of short periods of missing data using cubic spline and Fourier series approaches when the missing data gap is shorter than six hours. Cubic spline interpolation has been use to estimate the missing data when the gap is short in shootout II dataset. For most gaps that are much longer than six hours, the missing weather data were substituted by the measured data from the local weather station. This is not a principled way for missing data filling but extremely shortened the time of filling process.

Determination of Significant Wavelet Coefficients

Day Type Definitions

As described in Chapter IV, energy consumption data for the Zachry Engineering Center can be subdivided into 3 different types based on class schedule:

- Day type I: weekends and holidays

- Day type II: days other than type I and type III
- Day type III: all week days during spring and fall semesters except holidays

A daily energy use model will then be developed for each day type. The days with complete energy use data in each day type are used for model training and testing. In the current model, 2/3 of them are used for model training and 1/3 are used for model testing to determine whether the model is well-trained or not. The well-trained model is then used to predict removed energy use in the same day type.

All the predictors of the model are independent weather variables such as T_{out}, RH and I_{sol}. Wind speed V_{wind} is not a strong parameter related to building energy consumption in College Station, TX and has been excluded from predictor set.

Selection of Wavelet and Decomposition Level

The wavelet and wavelet decomposition level selection for the Zachry Engineering Center have been discussed in Chapter IV. For Energy Predictor Shootout II, relative humidity (RH), instead of $(\text{DPT-55})^{+}$, was used. The wavelets for T_{out}, RH and I_{sol} are db3, db1 and db1 respectively according to previous study. Daily average T_{out}, RH, I_{sol}, and detail coefficients of T_{out} at 5^{th} level and their combinations will be used as predictors of neural network models. The best predictors of daily energy use model will be determined.

Parameters of Neural Network Model

The basic neural network structure is a feed forward network containing an input layer, a hidden layer, and an output layer. The output layer is building daily energy use. The criterion of determining number of neurons in hidden layer is *CV*. An optimal number of neuron would generate close energy modeling *CV* for both training and testing datasets.

For the purpose of determining the most appropriate input set of the neural network, several combinations of daily average T_{out}, RH, I_{sol}, and detail coefficients of T_{out} at level $j = 5$ are tested. The combinations are listed in Table 7.1.

If the criterion for independent input selection is simply the best fit of the data, the model with more inputs would be considered. However, the *CV* does not seem to decrease significantly when increase the number of inputs. The most practical way is to use inputs as few as possible but still well enough to account for energy performance variations. In this comparison, up to three input variables are adopted.

Table 7.1. Daily Energy Model Inputs for the Zachry Building in Shootout II Comparison

Case	Neural Network Inputs
1	T_{out} at a_5
2	T_{out} at a_5, RH at a_5
3	T_{out} at a_5, I_{sol} at a_5
4	T_{out} at a_5, RH at a_5, I_{sol} at a_5
5	T_{out} at a_5, T_{out} at d_5
6	T_{out} at a_5, T_{out} at d_5, RH at a_5
7	T_{out} at a_5, T_{out} at d_5, I_{sol} at a_5

* T_{out} at a_5 is referred to as approximation coefficient of T_{out} at level $j = 5$ and T_{out} at d_5 is referred to as detail coefficient of T_{out} at level $j = 5$ according to Figure 3.4.

Significant Wavelet Coefficients for Cooling Energy Use

Based on the defined parameters of ANN, the Zachry Building cooling energy use data, and independent weather data, the neural network daily cooling energy use models are trained and tested. Table 7.2 has the coefficient of variation for each input case that measures model performance.

Table 7.2. Coefficient of Variation of Daily Cooling Energy Modeling for Different Input Cases in Shootout II Comparison

Case	CV_1 (%) **training dataset**	CV_2 (%) **testing dataset**	CV (%) **whole dataset**	**Error (%)** $(CV_1 - CV_2)/CV_1$
1	6.85	6.85	6.83	-0.03
2	5.84	6.11	5.91	-4.70
3	6.54	7.17	6.73	-9.55
4	5.77	5.80	5.74	-0.52
5	6.63	6.74	6.63	-1.66
6	5.62	5.66	5.60	-0.63
7	7.10	7.86	7.31	-10.59

A method to validate whether the network is well developed or not is to compare the energy modeling *CV* for training dataset and *CV* for testing dataset. The training dataset and the testing dataset are independent to each other, so a close *CV* means there is no over or under fitting of the data and the model is reliable to be applied to more general cases. Table 7.2 indicates an error of -0.63% for *CVs* between training and testing datasets for case 6. Simulated daily cooling energy use E_c for training dataset are shown in Figure 7.3 plotted against measured daily E_c. Figure 7.4 is for testing dataset of cooling energy use. The Case 6, where the model performance on training dataset and testing dataset are close enough, has the smallest *CV* compared to other cases. It can be concluded that daily average T_{out}, RH, and detail coefficients of T_{out} at level $j = 5$ are the best predictors for the Zachry building daily cooling energy use.

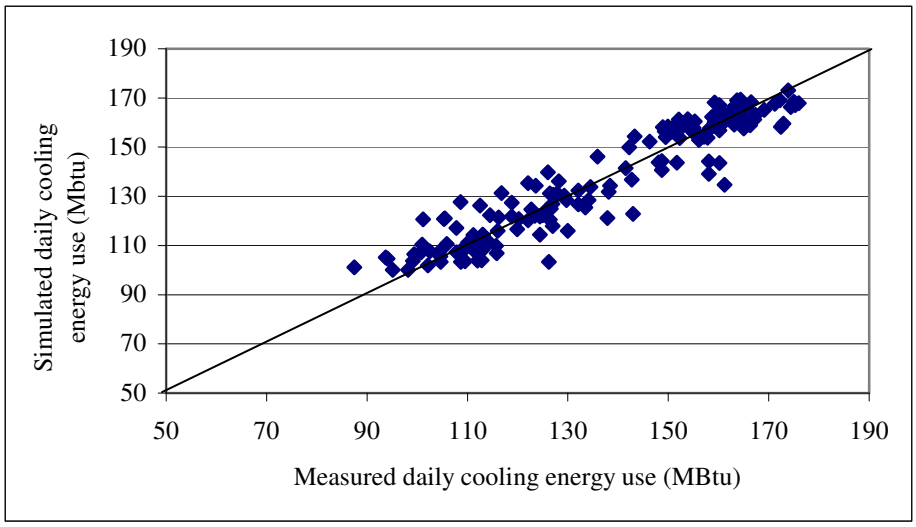


Figure 7.3. Measured and simulated E_c for training dataset.

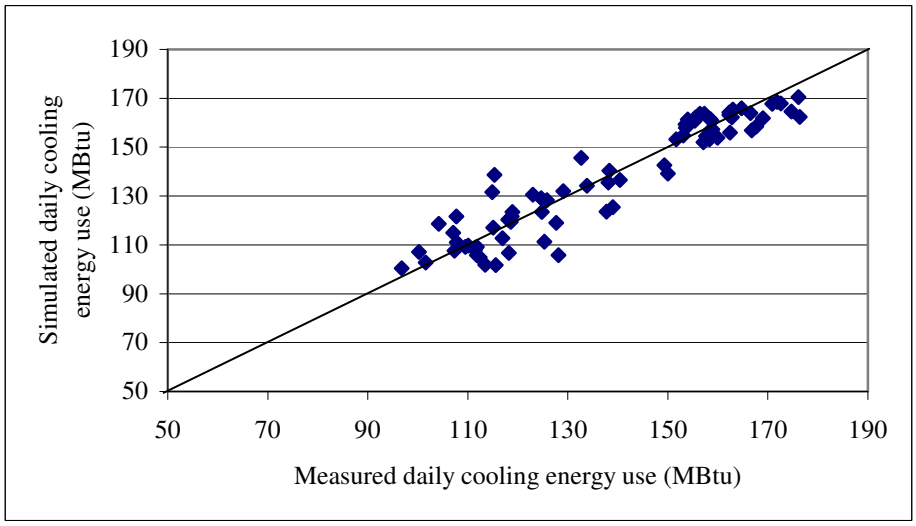


Figure 7.4. Measured and simulated E_c for testing dataset.

Significant Wavelet Coefficients for Heating Energy Use

Based on the defined parameters of ANN and the Zachry Building heating energy use data and independent weather data, the neural networks daily heating energy use model are trained and tested. Coefficients of variation for each case that measures model performance are listed in Table 7.3.

Table 7.3. Coefficients of Variation of Daily Heating Energy Modeling for Different Inputs in Shootout II Comparison

Case	**CV_1 (%) training dataset**	**CV_2 (%) testing dataset**	**CV (%) whole dataset**	**Error (%) $(CV_1-CV_2)/CV_1$**
1	21.02	24.90	22.31	-18.44
2	20.99	24.71	22.17	-17.75
3	20.52	25.46	22.15	-24.07
4	19.64	23.78	20.93	-21.08
5	18.74	20.42	19.22	-8.98
6	18.15	20.33	18.76	-12.05
7	18.64	20.98	19.30	-12.55

Table 7.3 indicates an error of -8.98% of modeling *CVs* between training and testing datasets for case 5. Simulated daily heating energy use E_h for training dataset are shown in Figure 7.5 plotted against measured daily E_h. Figure 7.6 is for testing dataset of heating energy use. The Case 5, where the model performance on training dataset and testing dataset are close enough, has the smallest CV compared to other cases. It can be

concluded that daily average T_{out} and detail coefficients of T_{out} at level $j = 0$ are the best predictors for the Zachry building daily heating energy use.

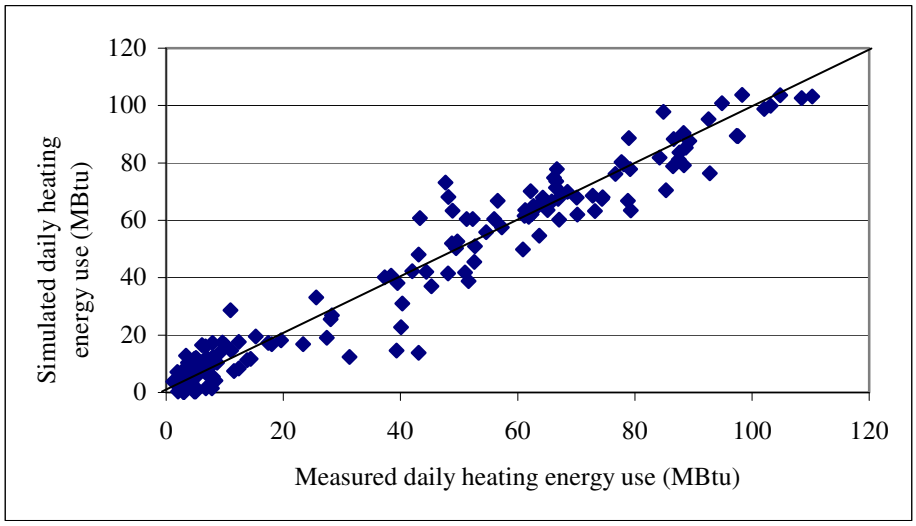


Figure 7.5. Measured and simulated E_h for training set.

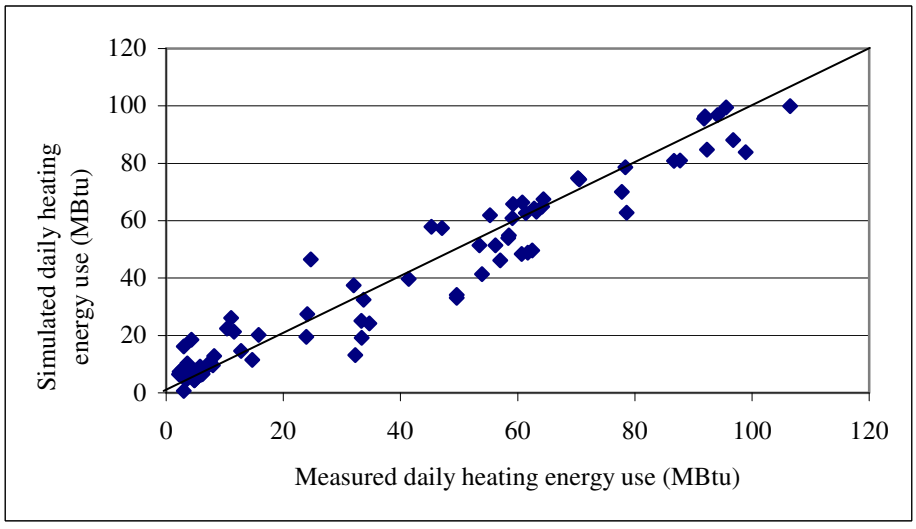


Figure 7.6. Measured and simulated E_h for testing set.

Neighborhood Classification

Weights Calculations

A multiple linear regression is taken with all significant wavelet coefficients against building daily energy use. The resulting coefficients of the regressors are weights of the significant wavelet coefficients. The significant wavelet coefficients for cooling energy use have been determined as daily average T_{out}, RH, and detail coefficients of T_{out} at level $j = 5$. The significant wavelet coefficients for heating energy use have been determined as daily average T_{out} and detail coefficients of T_{out} at level $j = 5$. The multiple linear regression yields weights of significant wavelet coefficients for cooling and heating energy use respectively as shown in Table 7.4 and 7.5.

Table 7.4. Weights of Significant Wavelet for Cooling Energy Use Model

Wavelet coefficients	T_{out} at a_5	T_{out} at d_5	RH at a_5
Weights	0.2486	0.0556	0.0394

Table 7.5. Weights of Significant Wavelet for Heating Energy Use Model

Wavelet coefficients	T_{out} at a_5	T_{out} at d_5
Weights	0.4128	0.3329

Determine Neighborhood

By feeding the significant wavelet coefficients and their corresponding weights of all the 330 days (except the first day because lagged variable data are used by energy use model) into SOM, the U-matrix representation of SOM suggested a classification of 3 neighborhoods. 3 is a reasonable number of neighborhood because each neighborhood would have enough days to represent climatic variation for energy model development. Table 7.6 lists the number of days in each neighborhood determined by SOM.

Table 7.6. Number of Days Classified in Each Neighborhood for the Zachry Building in Shootout II Comparison

	Nbhd 1	Nbhd 2	Nbhd 3	Total
Cooling Energy Use Model	66 days	126 days	138 days	330 days
Heating Energy Use Model	66 days	125 days	139 days	330 days

Hourly Energy Use Prediction Comparison with Shootout II

Neighborhoods have been classified by Self-organizing Map using significant wavelet coefficients that were determined by the ANN daily energy use models and the corresponding weights of wavelet coefficients that were determined by a simple multiple linear regression. In order to predict the removed energy use data, an ANN hourly energy model is developed for each neighborhood. Days with complete energy use data in each neighborhood are used for ANN model training. The well-trained model will be used to predict the removed energy use in the same neighborhood.

ANN Parameters

The hourly energy use ANN model is a feed-forward neural network model. For cooling energy use model, the input variables are current hour outdoor air temperature T_{out}, relative humidity RH, solar radiation I_{rad}, previous hour outdoor air temperature $T_{out}^{'}$, previous hour cooling energy use E_c, hour of the day and day type. The output variable is E_c. For heating energy use model, the input variables are current hour outdoor air temperature T_{out}, relative humidity RH, solar radiation I_{rad}, previous hour outdoor air temperature $T_{out}^{'}$, previous hour heating energy use E_h, hour of the day and day type. The output variable is E_h.

Energy Use Prediction and Comparison

Table 7.7 lists the resulting coefficient of variation *CV* and mean bias error *MBE* of ANN model performance on training dataset and testing dataset. The definition of *CV* is given in Chapter IV. The mean bias error *MBE* is defined as:

$$MBE(\%) = \frac{\dfrac{\sum_{i=1}^{n}(y_{pred,i} - y_{data,i})}{n-p}}{\overline{y}_{data}} \times 100$$

where, $y_{data,i}$ is measured energy use data for data point i; $y_{pred,i}$ is predicted energy use by the model for data point i; $\overline{y}_{data}$ is mean value of the measured energy use for the

dataset; n is number of data point in the dataset and p is total number of regressor variables in the model.

Cooling energy use modeling performances are shown in Figure 7.7, 7.8 and 7.9. Heating energy use modeling performances are shown in Figure 7.10, 7.11 and 7.12. ANN model training performances for cooling and heating energy use modeling are displayed in Figure 7.7 and Figure 7.10. These two figures indicate that ANN model simulated the training dataset very well. By comparing the energy modeling *CV*s between training dataset and testing dataset shown in Table 7.7, we can say that the models are well developed and are ready for removed data prediction. Figure 7.8 and Figure 7.11 have the simulated energy use for testing dataset. The prediction for 1864 missing hourly cooling energy use and 1871 missing hourly heating energy use against measured values are plotted in Figure 7.9 and Figure 7.12.

Table 7.7. Coefficient of Variation (*CV*) and Mean Bias Error (*MBE*) for Model Training and Testing

		Model training on training dataset	Model testing on testing dataset
Cooling energy use model	*CV* (%)	2.60	2.76
	MBE (%)	5.84e-6	0. 21
Heating energy use model	*CV* (%)	10.71	12.59
	MBE (%)	1.50e-7	1.28

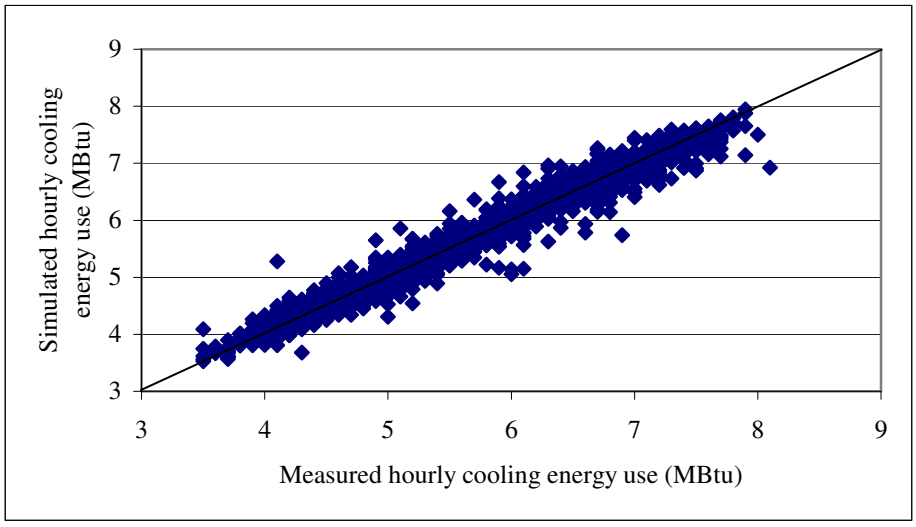


Figure 7.7. ANN model cooling energy use training output.

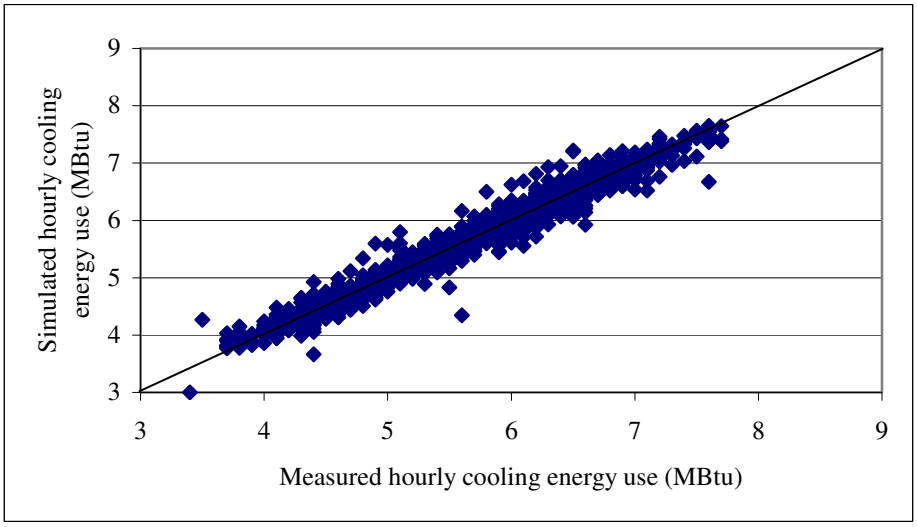


Figure 7.8. ANN model cooling energy use testing output.

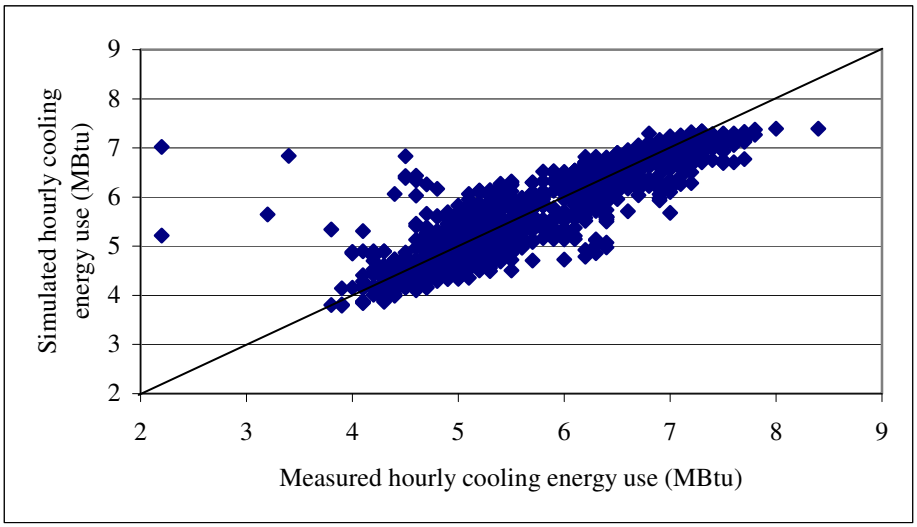


Figure 7.9. ANN model cooling energy use prediction output.

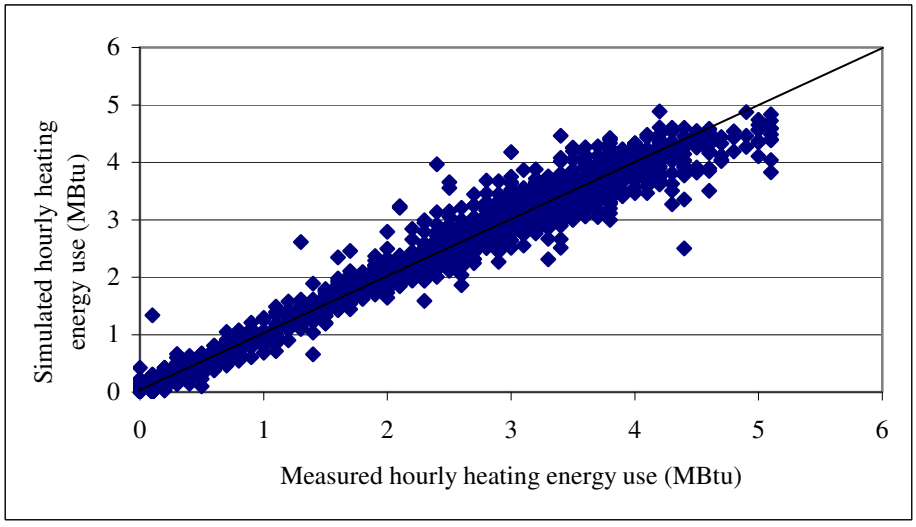


Figure 7.10. ANN model heating energy use training output.

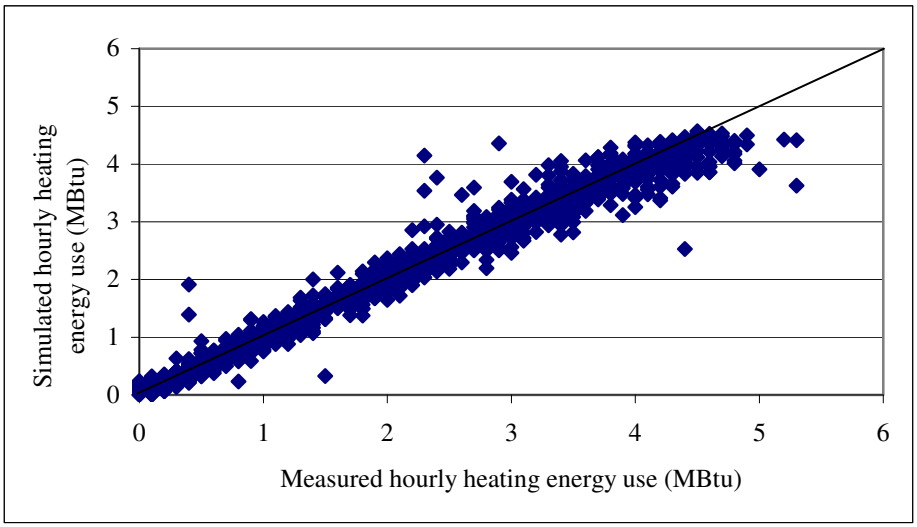


Figure 7.11. ANN model heating energy use testing output.

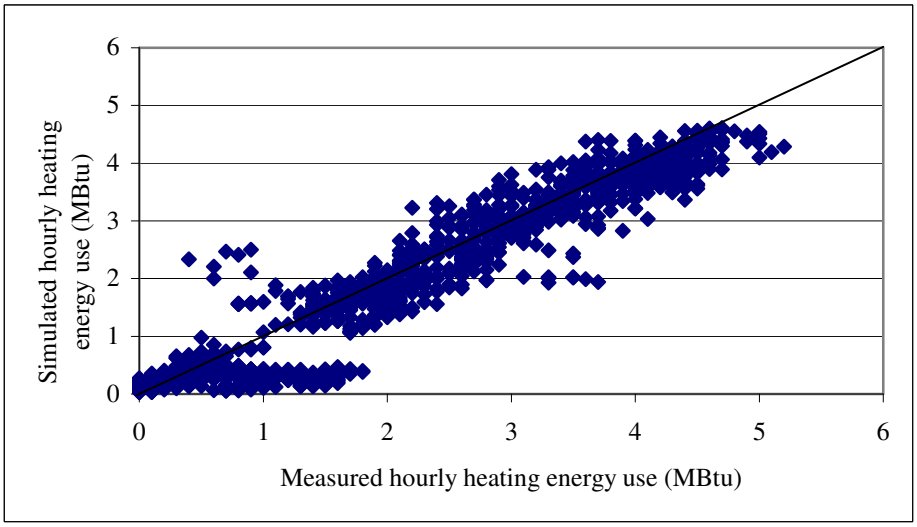


Figure 7.12. ANN model heating energy use prediction output.

The model statistics for missing data prediction are also calculated and tabulated with the competition winners as shown in Table 7.8. Table 7.9 illustrates analysis methods used by winners. Neighborhood based ANN model ranked best for cooling energy use prediction and ranked as the third best for heating energy prediction among the prediction models as shown in Table 7.8. It is worth pointing out that applying the neighborhood classification method to the Shootout II winning entries may increase their models' prediction accuracy.

Table 7.8. Comparison of the Neighborhood Based ANN Model Against the Competition Winning Entries (Haberl and Thamilseran, 1998)

		Winner 1	Winner 2	Winner 3	Winner 4	Neighborhood based ANN model
Cooling energy use	*CV* (%)	7.13	8.26	8.88	7.03	6.66
	MBE (%)	-0.89	-3.03	-3.42	-1.34	0.64%
Heating energy use	*CV* (%)	21.28	39.20	35.37	16.59	22.38
	MBE (%)	-3.10	-15.31	-10.98	-2.14	-3.8%

Table 7.9. The Methodologies of Winning Entries

	Contestants	Analysis Method
Winner 1	Chonan et al.	Bayesian nonlinear regression with multiple hyperparameters
Winner 2	Jang et al.	Feed-forward and autoassociative neural networks
Winner 3	Katipamula	Hourly weekday/weekend statistical multiple regression models
Winner 4	Dodier and Henze	Neural network models

Daily Energy Use Prediction Comparison with Change-Point Model

In the previous section, a comparison of neighborhood based ANN model to models of winning entries of shootout II was conducted. In this section, the comparison between neighborhood based linear regression and Chang-point model is performed. The choice of linear regression on each neighborhood instead of an ANN model is to make the two models more comparable. Shootout II data will be still used in the comparison.

Introduction of Change-Point Model

Heating and cooling energy consumption in commercial buildings may vary with outdoor air temperature throughout the entire range of outdoor air temperature encountered. Change-point models have the capability of capturing the variation for both heating and cooling energy use, and have had widespread use as baseline models for measuring energy savings (Haberl et al. 1998; DOE 1997).

A program of 4-parameter Change-point modeling has been developed. It uses a two-grid search to identify the best-fit change point (Kissock et al. 2003). The algorithm of linear regression over the change point is developed based on Vieth's method (Vieth 1989) and the RMSE is selected as the criterion to determine the best-fit change point for each iteration. The detailed code can be found in Appendix E.

Change-point Modeling for Shootout II Data

The performances of 4-paramenter change-point are shown in the following figures. Figure 7.13 depicts the development of change-point model using cooling energy use training dataset. The mathematic expression of this model is:

$$E_c = 161.68 - 1.6616(86.4 - T_{out})^+ + 0.002(T_{out} - 86.4)^+ \qquad (7.1)$$

where 86.4 $°F$ is the change point. 1.6616 is the left slope and 0.002 is the right slope. Figure 7.14 depicts the cooling energy use prediction using cooling energy change-point model when applied to testing dataset. *CV* of the prediction is 0.0761, and *MBE* is 0.0055.

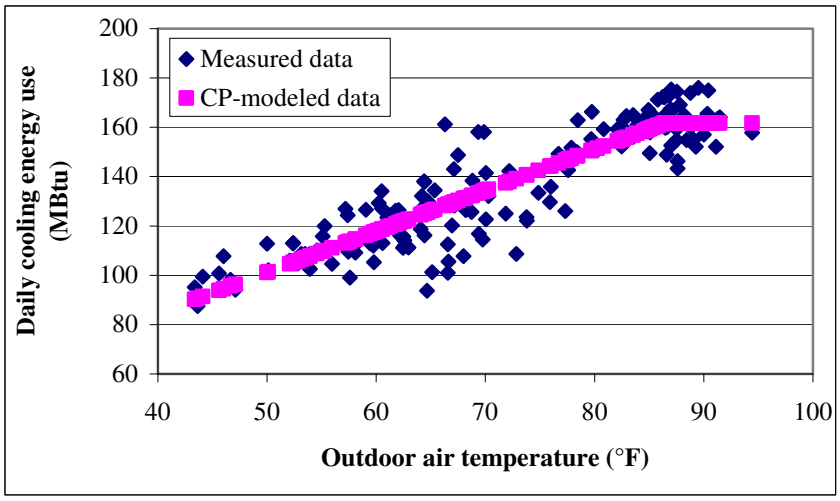


Figure 7.13. Cooling energy use CP model developed over training dataset.

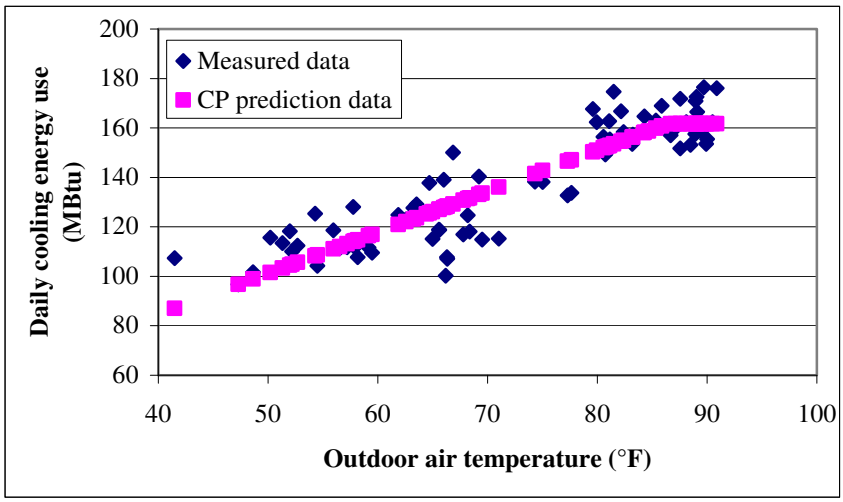


Figure 7.14. Cooling energy use prediction by CP model on testing dataset.

Figure 7.15 illustrates the development of change-point model using heating energy use training dataset. The mathematic expression of this model is:

$$E_h = 16.18 + 2.4842(81.06 - T_{out})^+ - 1.3719(T_{out} - 81.06)^+ \quad (7.2)$$

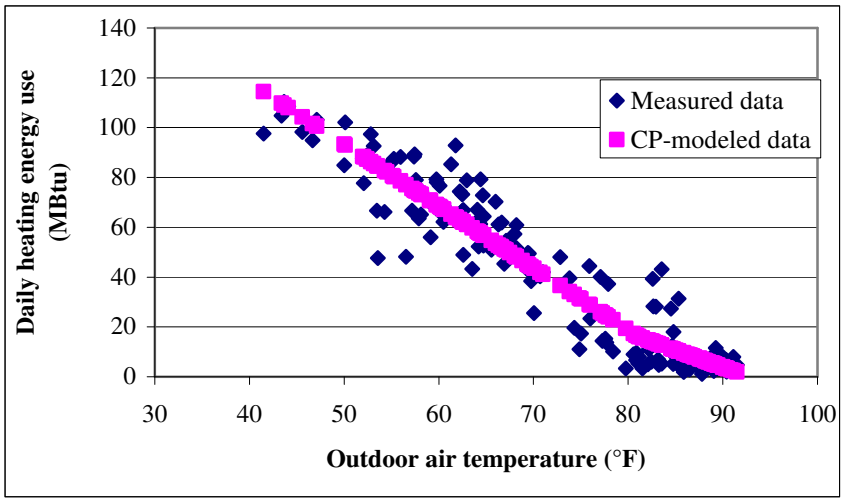


Figure 7.15. Heating energy use CP model developed over training dataset.

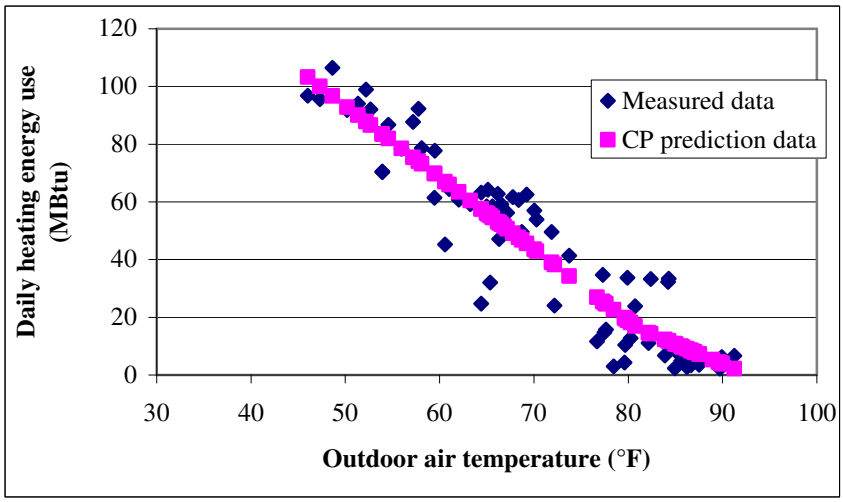
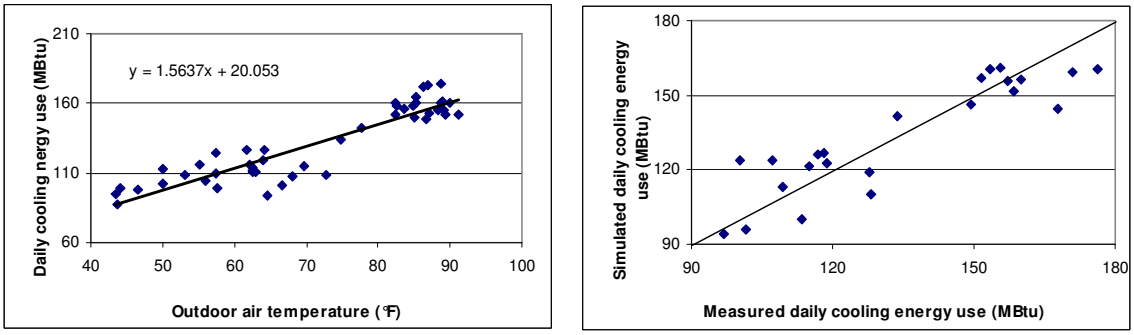


Figure 7.16. Heating energy use prediction by CP model on testing dataset.

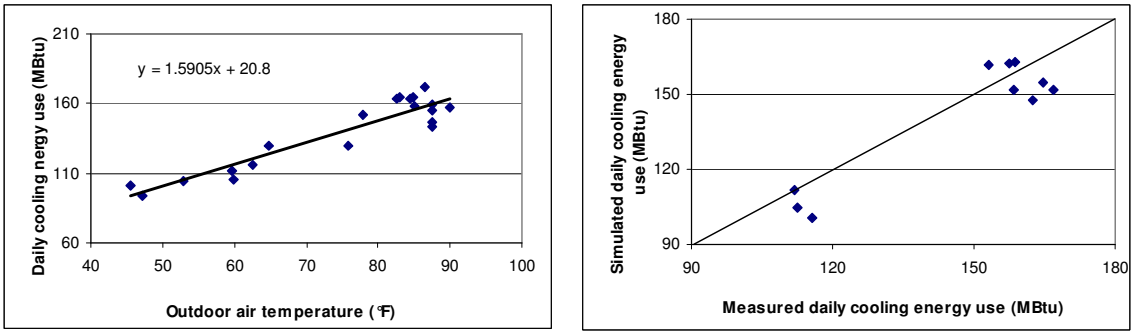
where 16.18 $°F$ is the change point. -2.4842 is the left slope and -1.3719 is the right slope. Figure 7.16 illustrates the heating energy use prediction using heating energy CP model when applied to testing dataset. *CV* of the energy use prediction is 0.2558, and *MBE* is 0.0084.

Neighborhood-based Linear Regression Model for Shootout II Data

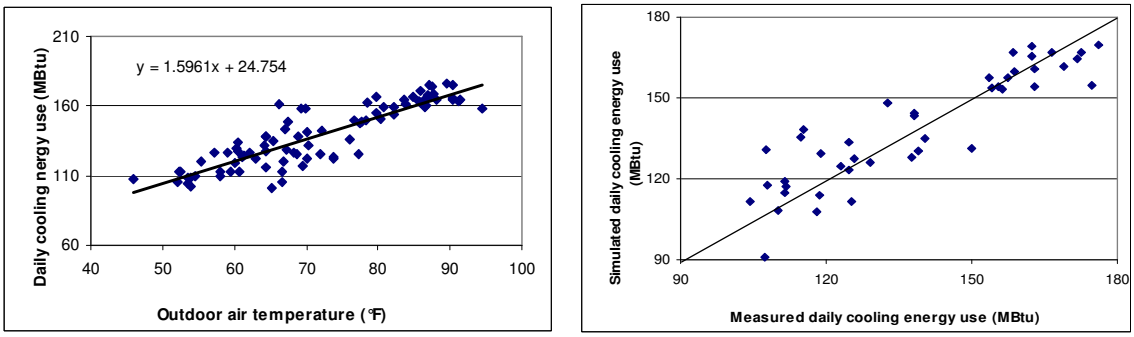
Base on the neighborhoods determined in the previous section, linear regression cooling and heating energy use models are developed for each of them. The training dataset in each neighborhood is used to develop the linear model and the testing dataset in each neighborhood is served to test the model performance. Figure 7.17 demonstrates the cooling energy use linear regression model and the corresponding prediction results for each of the neighborhood. Figure 7.18 demonstrates the heating energy use linear regression model and the corresponding prediction results for each of the neighborhood. *CV* of the cooling energy prediction is 0.0745, and *MBE* is -0.0047. *CV* of the heating energy prediction is 0.2481, and *MBE* is -0.0052.



(a) Linear regression model on neighborhood I.

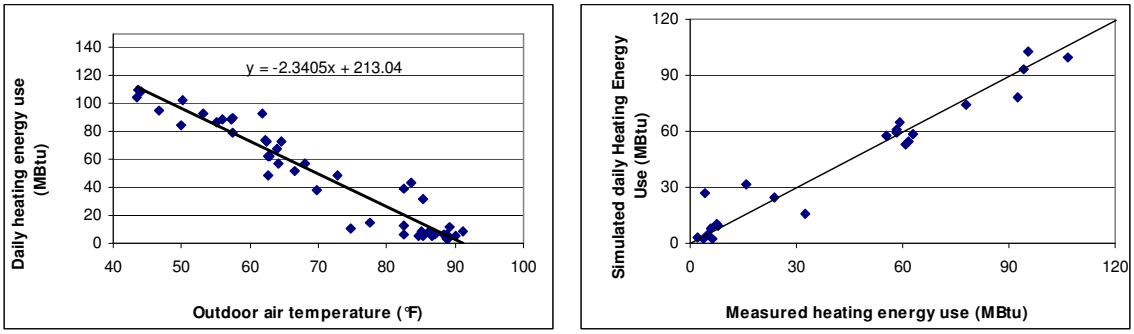


(b) Linear regression model on neighborhood II.

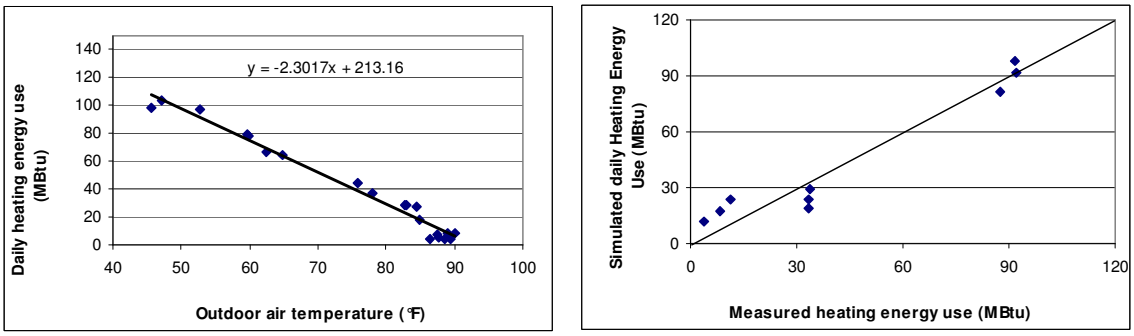


(c) Linear regression model on neighborhood III.

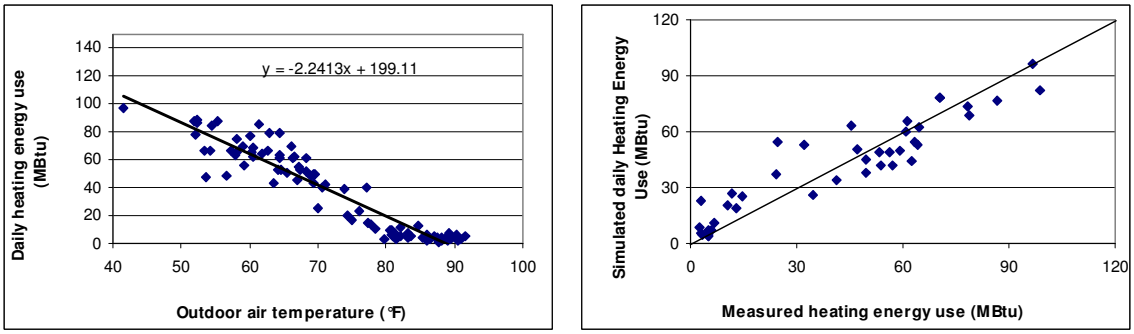
Figure 7.17. Neighborhood-based cooling energy use linear regression model (Left: model developed on training dataset; Right: simulated energy use on testing dataset).



(a) Linear regression model on neighborhood I.



(b) Linear regression model on neighborhood II.



(c) Linear regression model on neighborhood III.

Figure 7.18. Neighborhood-based heating energy use linear regression model (Left: model developed on training dataset; Right: simulated energy use on testing dataset) .

Statistic index of *CV* and *MBE* are listed in Table 7.10 for model comparison. Neighborhood based linear regression model did improve prediction accuracy compared to change-point model for the Zachry building but not significant. Change-point model is widely used daily energy use model with acceptable accuracy although outdoor air temperature is the only predictor in this model. Because the neighborhoods are determined by three weather components, using average outdoor air temperature as the only predictor may weaken the simulation capability of the neighborhood based models and shows little priority to the change-point model.

Table 7.10. Daily Energy Use Model Comparison with CP Model

		4-parameter change-point model	Neighborhood-based linear regression
Cooling energy use model	*CV* (%)	7.61	7.45
	MBE	0.0055	0.0047
Heating energy use model	*CV* (%)	25.58	24.81
	MBE	0.0084	-0.0052

Summary

Neighborhood base energy model has been applied to Shootout II data and compared with the winning entries for hourly energy use predictions. The cooling energy consumption prediction from the model showed better accuracy than all the wining entries in the competition, while the heating energy consumption prediction was found to have an average accuracy. It is worth mentioning that in this competition every contestant used different methods for missing data filling and hence used different

dataset to develop their models. A meticulously designed method for missing data filling would definitely improve modeling accuracy. Comparison to Change-point model for daily energy use prediction was also conducted using the same dataset. The vantage is not significant if using average outdoor air temperature as the only predictor. Using all the variables that defining neighborhoods as predictors would be more suitable for the models developed on the neighborhoods.

CHAPTER VIII

THE NEAREST NEIGHBORHOOD METHOD TO IMPROVE UNCERTAINTY ESTIMATES

Accurate estimation of uncertainty in energy use predictions from statistical models finds applications in a number of diverse areas of interest to building energy professionals. Some examples are in the determination of measured energy savings in monitoring and verification (M&V) projects, in automated fault detection, and in identifying improper building or equipment performance based on baseline model residual outliers. In this chapter, a general methodology for determining uncertainty in baseline models which is more realistic, and hence more robust and credible, than the statistical approaches currently used is introduced. The approach proposed in this chapter is to determine the uncertainty from "local" system behavior rather than from global statistical indices such as root mean square error and other measures as is the current practice. This is done using the non-parametric nearest neighborhood points approach which is well know in traditional statistics. The methodology is independent of the baseline model used (i.e., is applicable to any type of statistical model approach such as regression, time series, neural networks, …), and could be coded into a computer package that can be appended to existing M&V analysis programs. Two case study examples using daily building energy use data serve to illustrate the proposed methodology. The ultimate benefit of such an unambiguous, reliable and statistically defensible method is to lend more credibility to the determination of risk associated with

energy savings from energy efficiency projects, and thereby induce financial agencies to become more involved in “white tag” and allied certification programs.

Examples of energy prediction uncertainty analysis for a synthetic building and energy saving estimation uncertainty analysis will be studied and discussed in this chapter.

Energy Saving Estimation

Consider M&V programs used to estimate energy and cost savings resulting from energy conservation measures (ECMs). One of the three M&V options proposed by the IPMVP (2005) is to monitor specific end uses for a short period (of the order of weeks) before and after implementation of the ECM. The savings are estimated as the difference between pre- and post-ECM energy use normalized for factor such as weather, occupancy, etc. These ECMs could be any of a number of measures including modifications to the building internal loads or the HVAC system, replacement of existing equipment by more efficient equipment, or performing tune-ups or commissioning at various levels of detail. Some energy efficiency measures such as the installment of high efficiency lighting and motors are straightforward to evaluate from direct measurements. The benefits can be quantified and a robust market exists for these products. On the other hand, benefits from some measures such as building commissioning and optimal control of building HVAC systems are often not easy to determine through any simple direct measurements.

Requirements for Energy Saving Determination

The determination of savings from such ECMs requires that an accurate baseline model be identified (often based on statistical regression) in order to determine energy use had the ECMs not been implemented. The change-point models (Ruch and Claridge 1991; Kissock et al. 1998) or the multivariate linear models (Katipamula et al. 1998) usually serve as the baseline models. It also requires that uncertainty in the savings (determined as the difference between baseline model predicted and post-retrofit observed values) be ascertained properly. Professionals responsible for implementing M&V programs recognize the importance of determining this uncertainty as accurately or realistically as possible, and if feasible, minimizing it. The ability to determine this uncertainty provides both the energy professional and the building owner with a better sense of the risk involved in the stated energy savings estimate.

Energy Saving Estimation Using Neighborhood Based ANN Method

In Chapter VI, hourly energy use models were developed for each of the neighborhood classified based on the similarities of daily meteorological characteristics for pre-retrofit period. To estimate energy retrofit savings of a particular post-retrofit day, apply the characteristics of the day to the Self-organizing Map that was developed by training of the pre-retrofit days and find out the neighborhood it belongs to from the map. Use the hourly energy use ANN model corresponding to this neighborhood as the baseline energy model to calculate baseline energy use. Daily energy saving can then be determined by adding up energy savings of 24 hours.

Methodology of Uncertainty Analysis

A rather different approach for determining the uncertainty of savings estimates which relies on "local" model behavior as against global estimates such as the overall RMSE are proposed here. First consider the traditional method. A statistical model, such as a multiple linear regression (MLR) model or an artificial neural network model, of energy use is developed from pre-retrofit data. For a given post-retrofit day *j*, the model is used to determine what the energy use would have been in the absence of the retrofits. The difference between this and the measured energy use is the estimate of savings for the day:

$$E_{savings,j} = E_{pre,model,j} - E_{post,measured,j} \tag{8.1}$$

Following the industry-accepted approach (Kissock et al. 1998; Reddy et al. 1998; Reddy and Claridge 2000), the uncertainty in the savings $E_{savings,j}$ would be determined by measurement error in $E_{post,measured,j}$ and modeling error in $E_{pre,model,j}$. The latter error is usually more significant and its accuracy is hard to determine because the traditional uncertainty statistical formulae apply only to well-behaved model residuals with normal error distributions.

The central premise of the proposed methodology is that the uncertainty in this estimate is better characterized by identifying a certain number of days in the pre-retrofit period which closely match the specific values of the regressor set for the post-retrofit day *j*, and then determining the error distribution from this set of days. The distribution

of errors in $E_{post,measured,j}$ is specified by the set of errors associated with the points in its vicinity. However, the concept of vicinity requires a definition of the "distance" between days which is addressed below.

Assume that a statistical model with *p* regressor parameters has been identified from the pre-retrofit period based on daily variables. Any day, for example, day *j* can be represented as a point in this *p*-dimensional space. If data for a whole year are available, the days are represented by 365 points in this *p*-dimensional space. Associated with each point, or day *j*, is an error term which is identical to the model residual:

$$E_{error,j} = E_{meas,j} - E_{\text{model},j} \tag{8.2}$$

One could expect that energy use during days that are contiguous in the calendar will be akin to each other (with consideration given to day type: weekday, weekend, etc). This could be one criterion on which to select neighborhood points. Since one would expect the characteristics of the days relevant for energy use to be similar too, i.e., the regressor set of climatic and operational variables change slowly from one day to the next. Rather than to adopt this approach, the traditional, and the superior, approach has been to view the data as cross-sectional in nature, overlook the time series nature of the data and use statistical regression models. Commonly used regressors for daily energy use in a building are the average daily ambient temperature, solar radiation and humidity. A more exhaustive set would include in addition, not just the average daily values but the profiles during the day of ambient temperature, solar radiation and humidity as well.

It is therefore natural that statistical regression models for $E_{pre,model,j}$ are multivariate in nature.

The definition of the distance between two given days *i* and *j* specified by the set of regressor variables $X_{k,i}$ and $X_{k,j}$ is defined as:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} w_k^2 (X_{k,i} - X_{k,j})^2} = \sqrt{\sum_{k=1}^{p} (w_k X_{k,i} - w_k X_{k,j})^2} \tag{8.3}$$

where the weights w_k are given in terms of the derivative of energy *E* with respect to the regressors:

$$w_k = \left(\frac{\partial E_{pre,model}}{\partial X_k} \right) \tag{8.4}$$

The motivation for this definition of distance is that the weight given to a direction in the parameter space should be proportional to the rate of change of energy in that direction. Days that are at a given energy distance from a given day lie on an ellipsoid as illustrated in Figure 8.1.

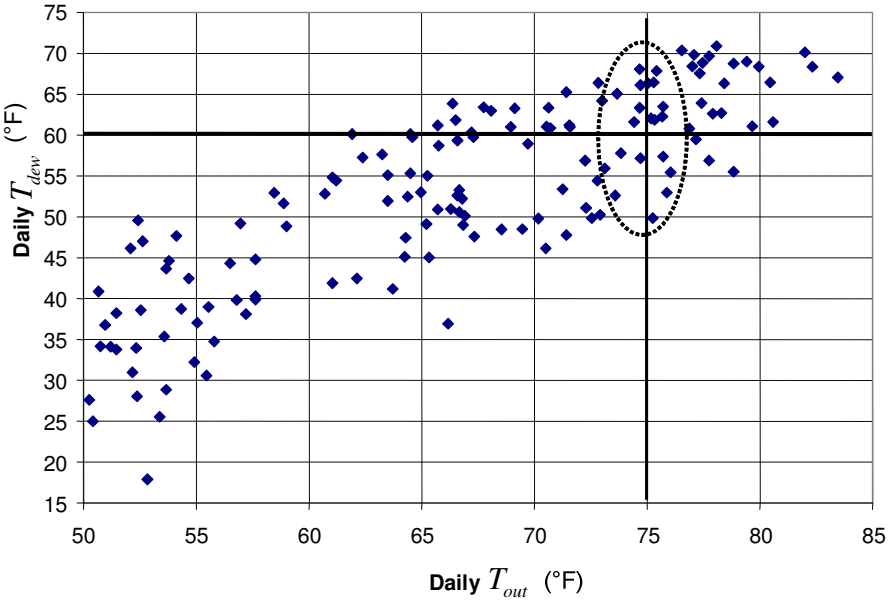


Figure 8.1. Illustration of the neighborhood concept for a baseline with two regressors (T_{out} and T_{dew}).

If the T_{out} variable has more "weight" than T_{dew} on the variation of the response variable, this would translate geometrically into an elliptic domain as shown in Figure 8.1. The data set of "neighborhood points" to the post datum point (75, 60) would consist of all points contained within the ellipse. Further, a given point within this ellipse may be assigned more "influence" the closer it is to the center of the ellipse.

Once a maximum distance is selected or if a pre-specified number of points are taken, an ellipsoid can be associated with each post-retrofit in the parameter space. Pre-retrofit days that lie inside this ellipsoid contribute to the determination of uncertainty in the estimation of the savings for this particular post-retrofit day. The overall size of the ellipsoid is determined by the requirements of making it as small as possible (so that

variations in the daily energy use are small) while having a sufficient number of pre-retrofit days within the ellipsoid. This may be a problem in sparsely populated regions, but one could argue that, for this very reason, the contribution of such regions to annual energy use is likely to be small.

The derivative in equation 8.4 can be determined by any accepted approach. However, the following calculation is suggested: Given the measured energy $E_{pre,i}$ for each pre-retrofit day, and the characteristics of each day, a method such as ANN is used to determine the functional dependence and obtain numerical derivatives. The recommendation for use of ANN models rather than the more-widely accepted procedures such as change-point or MLR models is because the ANN model approach provides a convenient way to capture non-linear functional discontinuities even though the physical significance is lost. Further, the residuals tend to be unbiased provided the ANN order, i.e., the number of nodes in the hidden layer, is fairly high.

Application of Uncertainty Analysis to Energy Use Prediction

The synthetic daily data of a large hospital in Newark that has been introduced in Chapter IV are used to illustrate the approach. The daily energy use model developed in Chapter IV has revealed that the daily average outdoor temperature and daily average effective dew point temperature are the best predictors for daily cooling energy use prediction. In this case, building internal electrical loads (E_{int}) due to lights and equipments are considered as predictor in order to check the energy use sensitivity to E_{int} . The hourly data has first been separated in weekdays and weekends and then

summed to represent daily values. Only the weekday data set consisting of 249 values were used to illustrate the concept of the proposed methodology. The correlation between T_{out} and ΔT_{dew} can be mitigated by a simple fitting between these two variables as introduced in Chapter V.

Consider the case to determine the uncertainty in the response variable corresponding to a set of operating conditions specified by T_{out}=75^0 F, ΔT_{dew}=5^0 F. The ANN model was used to numerically determine the gradients of these three regressors:

$$\frac{\partial E_c}{\partial (T_{out})} = 5.0685 \qquad \frac{\partial E_c}{\partial (\mathrm{Re}\, s T_{dew})} = 7.606 \qquad \frac{\partial E_c}{\partial E_{int}} = 0.050 \tag{8.5}$$

Note that though the sensitivity of the model to E_{int} appears small, its numerical value is high, and so its absolute impact on E_c is not negligible. However, this variable changes little from one weekday to the next over the year, typically by plus minus 20 kW from an average value of 598 kW. Consequently, in order to better illustrate the implementation of the proposed methodology, this variable has been dropped altogether.

The "distance" statistic for each of the 249 days in the synthetic data set has been computed following equation 8.3 and 8.4, and the data sorted by this statistic. The top 20

data points (with smallest distance) are shown in Table 8.1, as are the regressor values, the measured and predicted values, and their residuals. The last column assembles the "distance" variable. It may be noticed that this statistic varies from 1.78 to 23.06. In case, the 90% confidence intervals are to be determined, a distribution-free approach is to use the corresponding values of the 5^{th} and the 95^{th} percentiles of the residuals. Since there are 20 points, just reject the two extreme values of the residuals shown in Figure 8.2, which yields the 90% limits (-8.426 and 8.29) around the model predicted value of 233.88 MMBtu/day for the cooling energy use. In this case, the distribution is fairly symmetric, and one could report a local prediction value of (233.88 ± 8.36) at the 90% confidence level. If the traditional method of reporting uncertainty were to be adopted, the RMSE for the ANN model, found to be 5.7414 (or a $CV = 6.9\%$), would result in (± 9.44 MMBtu/day) at the 90% confidence level. Thus, in this case, there is some reduction in the uncertainty interval around the local prediction value. But more importantly, this estimate of uncertainty is more realistic and robust. Needless, to say, the advantage of this entire method is that even when the residuals are not normally distributed, the data itself can be used to ascertain statistical limits.

Table 8.1. Residual of the Twenty Nearest Neighborhood Points from A Reference Point of $T_{out} = 70\,°\text{F}$ and $\Delta T_{dew} = 5\,°\text{F}$

	T_{out}	$ResT_{dew}$	$E_{c,meas}$	$E_{c,model}$	E_{error}	distance
	(°F)	(°F)	(MMbtu/day)	(MMbtu/day)	(MMbtu/day)	
1	74.67	4.76	225.59	233.88	8.29	1.78
2	75.42	4.22	236.36	231.39	-4.97	4.47
3	77.08	5.45	240.21	248.19	7.98	7.85
4	76.54	6.23	251.99	252.94	0.95	8.62
5	77.00	4.08	239.01	238.55	-0.46	8.71
6	77.46	4.32	241.97	242.60	0.64	9.54
7	72.83	3.88	224.54	217.61	-6.93	9.82
8	77.75	5.00	240.63	247.13	6.50	9.86
9	75.04	2.88	221.23	223.84	2.61	11.40
10	75.29	2.85	224.36	222.07	-2.29	11.61
11	74.71	2.80	214.56	220.89	6.33	11.89
12	78.08	6.08	252.04	256.14	4.10	12.49
13	77.33	3.07	231.57	234.58	3.00	13.32
14	71.42	3.37	210.83	204.44	-6.39	15.53
15	78.83	3.61	238.85	242.55	3.70	15.63
16	73.67	2.19	200.35	209.02	8.66	15.87
17	79.42	3.60	250.10	241.68	-8.43	17.54
18	73.00	1.62	213.35	204.23	-9.12	19.55
19	79.96	2.74	240.66	239.73	-0.92	21.53
20	78.42	1.37	224.75	230.06	5.32	23.06

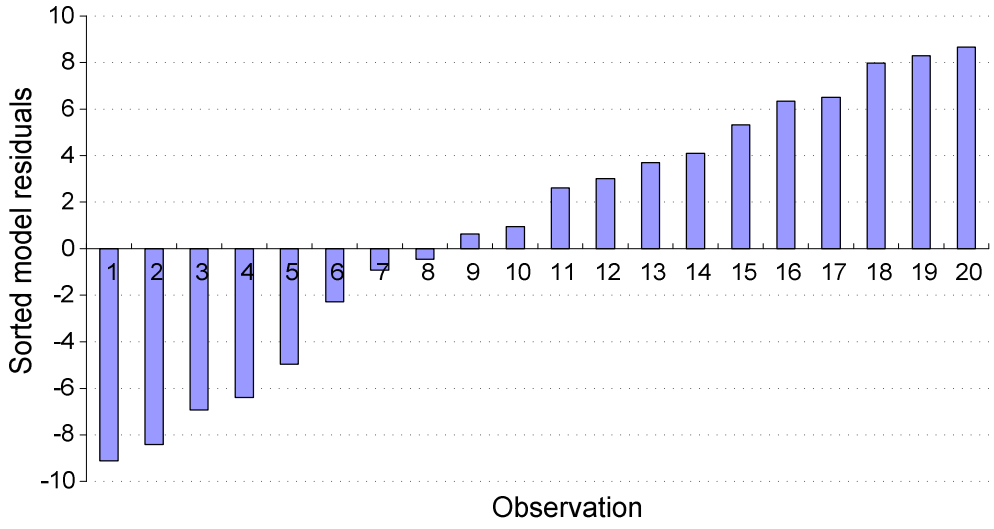


Figure 8.2. Plot of the sorted residuals for the 20 nearest neighborhood points to the reference point of $T_{out} = 70\,°\text{F}$ and $\Delta T_{dew} = 5\,°\text{F}$.

Application of Uncertainty Analysis to Energy Saving Estimation

The approach described above has been applied to data from the Zachry building which was used in Chapter IV. Daily energy model developed for the Zachry building in Chapter IV indicated that the daily average outdoor temperature, daily average effective dew point, and daily total global horizontal solar radiation are the best predictors for cooling energy use prediction. As in the previous section, daily total lights and equipment electrical energy usage (E_{int}) is also considered as predictor. A total of 91 days of data for the campus in session was used. As with the synthetic case study, $\text{Res}T_{dew}$ is used as the regressor variable. Energy saving is determined by equation 8.1.

Consider the problem of determining the savings for a given post-retrofit day, e.g., May 6, 1992. The measured E_c for that day was 94.5 MMBtu. The neural network

model of pre-retrofit energy use predicts that the energy use would have been 120.1 MMBtu without the retrofits. The savings estimate therefore is 25.6 MMBtu. The uncertainty distribution in the estimate of 120.1 MMBtu is determined from the prediction errors for the 20 pre-retrofit days closest to the post-retrofit day based on the regressor variables. The distance between days is defined through equation 8.3 and 8.4. The derivatives needed to evaluate the distance are very sensitive to overfitting. Small twists and turns in the response surface have significant effect on the derivative even thought the effect is not significant in the model fit. One approach is to make the model simpler by using fewer neurons in the model. Another is, just for the purposes of the derivatives, to use a simple linear regression model. The latter approached in this study. The resulting derivatives are:

$$\frac{\partial E_c}{\partial (T_{out})} = 1.27, \frac{\partial E_c}{\partial (\mathrm{Res}T_{dew})} = 2.40, \frac{\partial E_c}{\partial (E_{\mathrm{int}})} = -0.0058, \frac{\partial E_c}{\partial (I_{sol})} = 0.042$$

Daily internal gains vary little and therefore the corresponding derivative is not significant. The prediction error distribution is shown in Figure 8.3 ordered by distance. From these 20 days, the standard error is determined to be 6.4 MMBtu. If the measured post-retrofit energy use for the day has no errors, then the uncertainty in the savings estimate would be 6.4 MMBtu. Thus the final estimate of savings for the day is 25.6 ± 6.4 MMBtu. Had all 91 days been used the uncertainty estimate for this day (May 6) and for all the days would have been 4.3 MMBtu. The new method gives a more robust and realistic estimate.

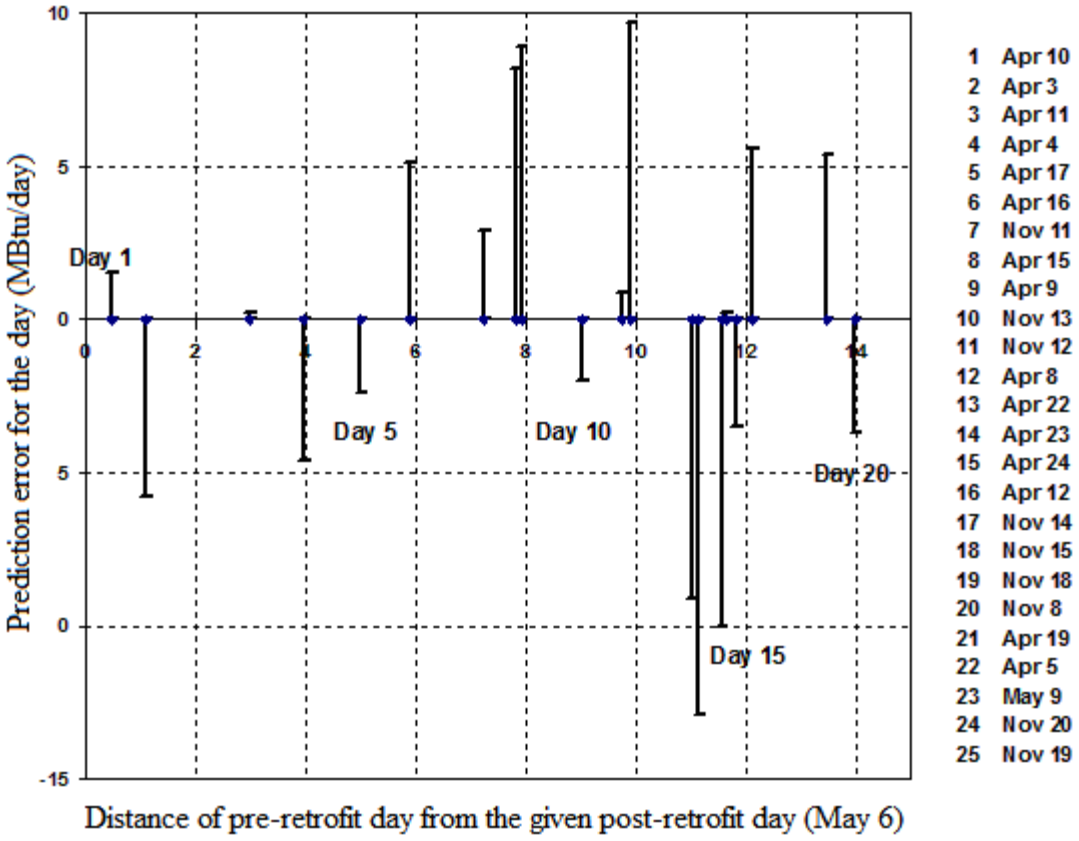


Figure 8.3. Residuals for the 20 pre-retrofit days in the data set closest to the selected post-retrofit day of May 6 (ordered by distance) .

Summary

This chapter elaborates a general methodology, using local system behavior, for determining uncertainty in baseline models which is more realistic, and hence more robust, than the current approaches. Rather than using global model goodness of fit measures, such as the RMSE and presume certain distributions for the residuals, this approach involves determining the uncertainty from "local" system behavior using the non-parametric nearest neighborhood points approach. The methodology is independent

of the baseline model used. Two case study illustrative examples using daily building energy use data, one based on synthetic data and the other on monitored data from two large commercial buildings, serve to illustrate the proposed methodology. The ultimate benefit of such an unambiguous, reliable and statistically defensible method is to lend more credibility to the determination of risk associated with energy savings from programs such as LEED certification, cap-and-trade programs for carbon dioxide emissions, and financing programs involving energy efficiency.

CHAPTER IX

SUMMARY AND FUTURE DIRECTIONS

Summary

Neighborhood-based daily and hourly energy use models have been developed in this dissertation. Days lead to similar building energy performance can be classified into the same neighborhood. A baseline model developed for that neighborhood can do prediction better than a global model because the days in the same neighborhood have similar meteorological characteristics. Wavelet analysis was employed for daily weather feature extraction. The daily meteorological features, also called significant wavelet coefficients from wavelet analysis, were utilized by Self-organizing Map to classify neighborhoods. This methodology was applied to The Great Energy Predictor Shootout II data. The comparisons to the Great energy prediction shootout II winning entries for hourly energy use prediction and to change-point model for daily energy use simulation were performed. The nearest neighborhood method to improve uncertainty estimates in building energy models was also studied.

Future Directions

Peak Load Prediction

Wavelet transforms are capable of doing both frequency domain and time domain analysis on a signal. Traditional Fourier transform is only suitable for frequency analysis. The ability of localization on time series makes discrete wavelet transform a

powerful tool in building peak load analysis. The relationship between peak load and wavelet coefficients in the vicinity of the peak load time location on time series would be interesting.

In most buildings, peak cooling loads occur in the afternoon or early evening. Unlike Fourier transform by which all the transformed frequency components are global to the daily climatic profile under study, we can find wavelet coefficients for just a short period around peak load time of the profile by DWT at an adequate scale and location. The daily peak loads could then be represented by both global wavelet coefficients (average OAT, etc.) and localized wavelet coefficients. Figure 9.1 illustrates daily temperature and cooling energy use profile for the Zachry building in College Station, TX in July 28, 1990.

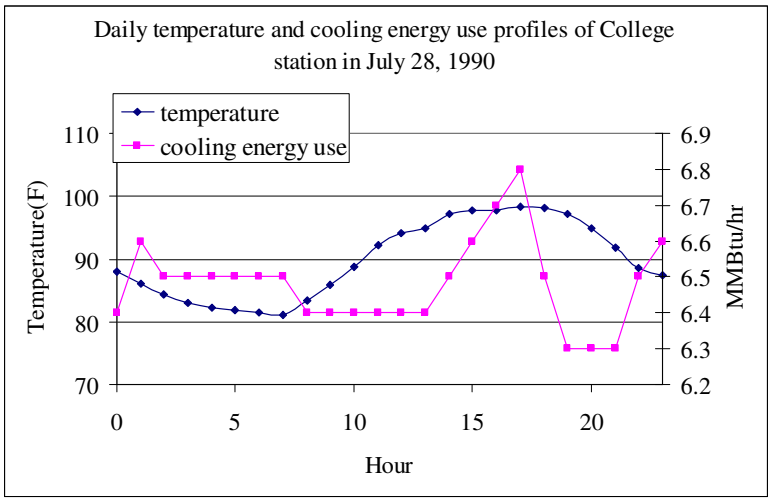


Figure 9.1. Daily temperature and cooling energy use profile for the Zachry Building.

From the figure, the peak load occurred at around 5pm. Wavelet coefficients corresponding to 5pm at a certain scale (frequency) may be good predictors for peak load. A peak load model can be developed using peak load wavelet coefficients of T_{out}, T_{dew} and, I_{sol} etc. as regressors in addition to daily average T_{out}, T_{dew}, and I_{sol} etc.

Solar Heat Gain

Cooling load for many buildings are solar driven, which means the solar radiation accounts for a large portion of the total cooling load. For these buildings, solar radiation, e.g. the usually used global horizontal radiation, would not be suitable to represent solar heat gain of the buildings in energy modeling. For statistical regression models, it is hard to establish an explicit relationship between the solar radiation and cooling load. Neural network model can better elaborate this relationship by adjusting weights of neurons and their bias terms. But we still suggest using solar gain directly to replace solar radiation.

An initiative study has been performed by using solar gain instead of solar radiation in the synthetic building energy simulation. We used DOE2.1e to calculate building cooling load from the solar radiation only by balancing the other parts of heat transfer from the environment. By setting building cooling load from solar gain as one of the model inputs, the neural network model demonstrated a much better simulation results. Due to the complexity of DOE2 programming, a simplified solar gain algorithm combined with the data-driven energy baseline model may be an interesting future direction.

REFERENCES

Akbari, H., K. Heinemeier, P. LeConiac, and D. Flora. 1988. An algorithm to disaggregate commercial whole-building hourly electrical load into end uses. *Proceedings of the ACEEE 1988 Summer Study of Environmental Efficiency in Buildings* 10:14-26.

ASHRAE. 2002. *Guideline 14-2002, Measurement of Energy and Demand Savings*. Atlanta: American Society of Heating, Refrigeration, and Air-conditioning Engineers, Inc.

ASHRAE. 2005. *2005 ASHRAE Handbook-Fundamentals*. Atlanta: American Society of Heating, Refrigeration, and Air-conditioning Engineers, Inc.

Addison, P.S. 2002. *The Illustrated Wavelet Transform Handbook*. Bristol and Philadelphia: Institute of Physics Publishing.

Boonyatikarn, S. 1982. Impact of building envelopes on energy consumption and energy design guidelines. *Proceedings of the ASHRAE/DOE Conference: Thermal Performance of the Exterior Envelope of Buildings* II 469-75.

Brislawn, C. 1995. Fingerprints go digital. *Notices of the AMS* 42(11): 1278-83.

Building Systems Laboratory (BSL). 1999. *BLAST 3.0 Users Manual*. Department of Mechanical and Industrial Engineering, Building Systems Laboratory, University of Illinois, Urbana-Champaign, IL.

Chonan, Y., K. Nishida, and T. Matsumoto. 1996. Great energy predictor shootout II-A Bayesian nonlinear regression with multiple hyperparameters. *ASHRAE Transactions* 102(2): 404-11.

Daneshdoost, M., M. Lotfalian, G. Bumroonggit, and J. Ngoy. 1998. Neural network with fuzzy set-based classification for short-term load forecasting. *IEEE Transactions on Power Systems* 13(4): 1386–91.

Davis, J.M., B.K. Eder, D. Nychka and Q. Yang. 1998. Modeling the effects of meteorology on ozone in Houston using cluster analysis and generalized additive models. *Atmospheric Environment* 32:2505-20.

Debnath, L. 2002. *Wavelet Transforms and Their Applications*. Boston: Birkhäuser.

Dhar, A., 1995, Development of Fourier series and artificial neural network approaches to model hourly energy use in commercial buildings. PhD dissertation, Texas A&M University, College Station, TX.

Dodier, R. and G. Henze. 1996. Statistical analysis of neural network as applied to building energy prediction. *Proceedings of the ASME ISEC, San Antonio, TX*, pp. 495–506.

DOE. 1997, *International Performance Measurement and Verification Protocol*. U.S. Department of Energy, Washington, DC.

Drezga, I. and S. Rahman. 1999. Short-term load forecasting with local ANN predictors. *IEEE Transactions Power Systems* 14(3):844–50.

Eder, B.K., J.M. Davis and P. Bloomfield. 1994. An automated classification scheme designed to better elucidate the dependence of ozone on meteorology. *Journal of Applied Meteorology* 33:1182-99.

EPA. Clean energy policy maps. http://www.epa.gov/cleanenergy/energy-programs/state-and-local/policy-maps.html. Accessed in October, 2008.

Eto, J. 1998. On using degree-days to account for the effects of weather on annual energy use in office buildings. *Energy and Buildings* 12(2):113-27.

Feinauer, D. 2007. Methodologies for developing energy consumption baseline for TAMU buildings. Master of Engineering Project, Texas A&M University, College Station, TX.

Fels, M. 1986a. PRISM: An introduction. *Energy and Buildings* 9:5-18.

Fels, M. 1986b. Special issue devoted to measuring energy savings: The scorekeeping approach. *Energy and Buildings* 9(1&2).

Fels, M. and M. Goldberg. 1986. Using the scorekeeping approach to monitor aggregate energy conservation. *Energy and Buildings* 9:161-68.

FEMP (Federal Energy Management Program). 2000. *Measurement and verification (M&V) Guidelines for Federal Energy Projects*. Washington, DC: U.S. Department of Energy.

Fowlkes, C. W. 1985. Snapshot: A Short-Term Building Energy Monitoring Methodology. Bozeman, MT: Fowlkes Engineering.

Goldberg, M. 1982. A geometric approach to nondifferentiable regression models as related to methods for assessing residential energy conservation. PhD dissertation, Princeton, NJ: Princeton University.

Gouda, M.M., S. Danaher, and C.P. Underwood. 2006. Quasi-adaptive fuzzy heating control of solar buildings. *Building and Environment* 41:1881-91.

Grossman, A. and J. Morlet. 1984. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM Journal of Mathematical Analysis* 15:723-36.

Haar A. 1910. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen* 69:331-71.

Haberl, J., C. Culp, and D. E. Claridge. 2005. ASHRAE's Guideline 14-2002 for measurement of energy and demand savings: How to determine what was really saved by the retrofit. Energy Systems Laboratory, Texas A&M University.

Haberl, J. and D. E. Claridge. 1987. An expert system for building energy consumption analysis: Prototype Results. *ASHRAE Transactions* 93(1):979-87.

Haberl, J. and E. Vajda. 1988. Use of metered data analysis to improve building operation and maintenance: Early results from two federal complexes. *Proceedings of the ACEEE 1988 Summer Study on Energy Efficient Buildings, Pacific Grove, CA, August,* pp. 3.98-3.111.

Haberl, J. and P. Komer. 1990. Improving commercial building energy audits: how daily and hourly data can help. *ASHRAE Journal* 32(9):26-36.

Haberl, J. and S. Thamilseran. 1996. The great energy predictor shootout II: Measuring retrofit savings – overview and discussion of results. *ASHRAE Transactions* 102(2): 419-35.

Haberl, J., S. Thamilseran, T.A. Reddy, D.E. Claridge, D. O'Neal, and D. Turner. 1998. Baseline calculations for measurement and verification of energy and demand savings in a revolving load program in Texas. *ASHRAE Transactions* 104(2):841-58.

Haberl, J., A. Sreshthaputra, D. E. Claridge, and J. Kissock. 2003. Inverse Model Toolkit: Application and testing. *ASHRAE Transactions-Research* 109(2):435–48.

Hadley, D. and S. Tomich. 1986. Multivariate statistical assessment of meteorological influences on residential space heating. *Proceedings of the ACEEE 1986 Summer Study on Energy Efficiency in Buildings* pp. 9.132-9.145.

Hadley, D. 1993. Daily variations in HVAC system electrical energy consumption in response to different weather condition. *Energy and Buildings* 19(3):235-47.

Hagan, M.T., H.B. Demuth, and M. Beale. 1996. *Neural Network Design*. Boston: PWS Publishing Company.

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

IPMVP (*International Performance Measurement Verification Protocol*). 2007. Washington DC: U.S. Department of Energy.

Jang, K., E. Bartlett, and R. Nelson. 1996. Measuring retrofit energy saving using autoassociative neural networks. *ASHRAE Transactions* 102(2): 412-18.

Katipamula, S. 1996. Great energy predictor shootout II: Modeling energy use in large commercial buildings. *ASHRAE Transactions* 102(2): 397-404.

Katipamula, S., T.A. Reddy, and D.E. Claridge. 1998. Multivariate regression modeling. *Journal of Solar Energy Engineering* 120:177-84.

Kermanshahi, B.S., C.H. Poskar, G. Swift, P. Mclaren, W. Pedrycz, W. Buhr, A. Silk. 1993. Artificial neural network for forecasting daily loads of a Canadian electric utility. *Proceeding of IEEE Second International Forum on Application of Neural Networks to Power Systems (ANNPS'93), Yokohama, Japan*, pp. 302-07.

Kissock, K., J. Haberl, and D.E. Claridge. 2003. Inverse modeling toolkit: Numerical algorithms. *ASHRAE Transactions-Research* 109(2):425–34.

Kissock, K. 1993. A methodology to measure energy savings in commercial buildings. PhD dissertation, Mechanical Engineering Department, Texas A&M University, College Station, TX.

Kissock, K., T.A. Reddy, J. Haberl, and D.E. Claridge. 1993. EModel: A new tool for analyzing building energy use data. College Station, TX: Energy Systems Lab, ESL-PA-93/03-04.

Kissock, K. and M. Fels. 1995. An assessment of PRISM's reliability for commercial buildings. *Proceedings of the National Energy Program Evaluation Conference*, *Chicago, IL,* pp.197-202.

Kissock, K., T.A. Reddy, and D.E. Claridge. 1998. Ambient temperature regression analysis for estimating retrofit savings in commercial buildings. *ASME Journal of Solar Energy Engineering* 120(3):168-76.

Kohonen, T. 1981. Automatic formation of topological maps of patterns in a self-organizing system. *Proceedings of 2SCIA, Scand. Conference on Image Analysis. Helsinki, Finland*, pp. 214-20.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43(1):59-69.

Kohonen, T. 2001. *Self-Organizing Maps*. New York: Springer.

Kohonen, T., K. Mäkisara, and T. Saramäki. 1984. Phonotopic maps - insightful representation of phonological features for speech recognition. *Proceedings of 7ICPR, International Conference on Pattern Recognition*, *Los Alamitos, CA*, pp. 182-85.

Kohonen, T. and O. Simula. 1996. Engineering applications of the Self-Organizing Map. *Proceedings of the IEEE* 84(10): 1358-84.

Kreider, J. F. and X.A. Wang. 1991. Artificial neural networks demonstration for automated generation of energy use predictors for commercial buildings *ASHRAE Transactions* 97(1):775-79.

Kreider, J. and J. Haberl. 1994. Predicting hourly building energy use: The great energy predictor shootout – Overview and discussion of results. *ASHRAE Transactions* 100(2): 1104–18.

Kreider, J., D. Claridge, P. Curtiss, R. Dodier, J. Haberl, and M. Krarti. 1995. Building energy use prediction and system identification using recurrent neural networks. *ASME Transactions Journal of Solar Energy Engineering* 117(3):161-66.

Kutscher, C.F. 2007. *Tackling Climate Change in the U.S. - Potential Carbon Emissions Reductions from Energy Efficiency and Renewable Energy by 2030*. Washington DC: American Solar Energy Society.

LBL. 1993. *DOE-2 Program: Version 2.1E*. Berkeley, CA: Lawrence Berkeley Laboratory.

Leslie, N. P., G.A. Aveta, and B.J. Sliwinski. 1986. Regression based process energy analysis system. *ASHRAE Transactions* 92 (1):93-102.

Lu, C.N., H.T. Wu, and S. Vemuri. 1993. Neural network based short term load forecasting. *IEEE Transactions on Power Systems* 8(1):336-42.

Lu, H., J. Hsieh, and T. Chang. 2006. Prediction of daily maximum ozone concentrations from meteorological conditions using a two-stage neural network. *Atmospheric Research* 81:124–39.

MacDonald, J.M. and D.M. Wasserman. 1989. *Investigation of Metered Data Analysis Methods for Commercial and Related Buildings*. Oak Ridge, TN: Oak Ridge National Laboratory Report ORNL/CON-279.

MacKay, D. 1994. Bayesian non-linear modeling for the prediction competition. *ASHRAE Transactions* 100(2):1053-62.

Mallat, S. 1988. Multiresolution representations and wavelets. PhD dissertation, University of Pennsylvania, Philadelphia.

Mandal, P., T. Senjyu, N. Urasaki, and T. Funabashi. 2006. A neural network based several-hour-ahead electric load forecasting using similar days approach. *International Journal of Electrical Power & Energy Systems* 28(6): 367-73.

Mangiameli, P., S.K. Chen, and D. West. 1996. A comparison of SOM of neural network and hierarchical methods. *European Journal of Operational Research* 93:402-17.

Mazzucchi, R. P. 1986. The project on restaurant energy performance end-use monitoring and analysis. *ASHRAE Transactions* 92(2): 328-52.

Maor, I. and T.A. Reddy. 2008. *Near-Optimal Scheduling Control of Combined Heat and Power Systems for Buildings*. Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Miller, R. and J. Seem. 1991. Comparison of artificial neural networks with traditional methods of predicting return from night setback. *ASHRAE Transactions* 97(2):500-08.

Mujica, L., J. Vehí, and J. Rodellar. 2005. A hybrid system combining Self Organizing Maps with case based reasoning in structural assessment. *Artificial Intelligence and Development*. 173-180. Amsterdam: ISO Press.

National Renewable Energy Laboratory (NREL). 1995. *User's Manual for TMY2s (Typical Meteorological Years)*. NREL/SP-463-7668. Golden, CO: NREL.

Neter, J., W. Wasserman, and M. Kutner. 1989. *Applied Linear Regression Models*. Boston: Irwin, Inc.

Rabl, A. 1988. Parameter estimation in buildings: Methods for dynamic analysis of measured Energy Use. *Journal of Solar Energy Engineering* 110:52-66.

Rabl, A., L. Norford, and J. Spadaro. 1986. Steady state models for analysis of commercial building energy data. *Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings, Santa Cruz, CA, August*, pp. 9.239-9.261.

Rader, N. and R.B. Norgaard. 1996. Efficiency and sustainability in restructured electricity markets: The renewables portfolio standard. *The Electricity Journal* 9:37–49.

Reddy, T. A.1989. Application of dynamic building inverse models to three occupied residences monitored non-intrusively. *Proceedings of the ASHRAE/DOE/BETCC/CIBSE Conference: Thermal Performance of the Exterior Envelopes of Buildings* IV: 654-71.

Reddy, T. A. and D.E. Claridge. 1994. Using synthetic data to evaluate multiple regression and principal component analyses for statistical modeling of daily building energy consumption. *Energy and Buildings* 21:35-44.

Reddy, T.A., N.F. Saman, D.E. Claridge, J.S. Haberl, W.D. Turner, and A. Chalifoux. 1997. Baselining methodology for facility level monthly energy use – Part 1: Theoretical aspects. *ASHRAE Transactions* 103(2):336-47.

Reddy, T.A., K. Kissock, and K. Ruch. 1998. Uncertainty in baseline regression modeling and in determination of retrofit savings. *ASME Journal of Solar Energy Engineering* 120:185-92.

Reddy, T.A. and D. Claridge. 2000. Uncertainty of measured energy savings from statistical baseline models. *International Journal of HVAC&R Research* 6(1):3-20.

Ruch, D. and D. Claridge. 1991. A four parameter change-point model for predicting energy consumption in commercial buildings. *Proceedings of the ASME International Solar Energy Conference, Reno, Nevada*, pp.433-40.

Ruch, D. and D. Claridge. 1992. A four parameter change-point model for predicting energy consumption in commercial buildings. *ASME Journal of Solar Energy Engineering* 114(2):77 -83.

Ruch, D., L. Chen, J. Haberl, and D. Claridge. 1993. A change point principal component analysis (CP/PCA) method for predicting energy usage in commercial buildings: The PCA model. *ASME Journal of Solar Energy Engineering* 115(2):77-84.

Ruch, D., J. Kissock, and T.A. Reddy. 1999. Prediction uncertainty of linear building energy use models with autocorrelated residuals. *ASME Journal of Solar Energy Engineering* 121(1):63-68.

Schrock, D. and D. Claridge. 1989. Predicting energy usage in a supermarket. *Proceedings of the Sixth Symposium on Improving Building Systems in Hot and Humid Climates. Dallas, TX. October*. pp. 44-54.

Smith, S. 2003. *Digital Signal Processing: A Practical Guide for Engineers and Scientists*. Burlington, MA: Elsevier Science.

Sonderegger, R.A. 1998. Baseline model for utility bill analysis using both weather and non-weather-related variables. *ASHRAE Transactions* 104(2):859-70.

Stoer J. and R. Burlirsch. 1996. *Introduction to Numerical Analysis*. 2nd edition. New York: Springer.

Subbarao, K. 2001. Method and apparatus for improving building energy simulations. U.S. Patent Number 6,134,511.

Thamilseran, S. and J. Haberl. 1995. A bin method for calculating energy conservation retrofit savings in commercial buildings. *Solar Engineering* 1: 111-23.

Thamilseran, S., 1999, An inverse bin methodology to measure the savings from energy conservation retrofits in commercial buildings. PhD dissertation, Texas A&M University, College Station, TX.

Ultsch, A. and H.P. Siemon. 1990. Kohonen's self organizing feature maps for exploratory data analysis. *Proceedings of the International Neural Network Conference, Dordrecht, Netherlands.* pp. 305-08.

UIUC (University of Illinois at Urbana Champaign) and LBNL (Lawrence Berkeley National Laboratory). 2005. *EnergyPlus Engineering Reference: The Reference to EnergyPlus Calculations*. Washington DC: U.S. Department of Energy.

Usevitch, B. 2001. A tutorial on modern lossy wavelet image compression: Foundations of JPEG 2000. *IEEE Signal Processing Magazine* 18:22-35.

Vieth, E. 1989. Fitting piecewise linear regression functions to biological responses. *Journal of Applied Physiology* 67(1):390-96.

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas. 1999. Self-Organizing Map in Matlab: the SOM Toolbox. *Proceedings of the Matlab DSP Conference, Finland, November,* pp. 35-40.

APPENDIX A: BUILDING INFORMATION

Table A1. Summary of Large Hospital Building Description

General	
Floor Area (ft^2)	315,000
Above Grade Floors	7
Below Grade Floors	0
% Conditioned and Lit	100
Geometry	
Footprint Shape	Rectangular (300' X 150')
Zoning (1st thru 6th floor)	4 Perimeter/1 Interior
Zoning (7th floor)	4 Perimeter/1 Interior
Perimeter Depth (feet)	15
Floor to floor height (ft)	13
Floor to ceiling height (ft)	9
Envelope	
Roof	Massive, R-19
Walls	CMU Grouted, 2" Insulation, U=0.1 $Btu/(h \cdot ft^2 \cdot F)$
Foundation	Slab, U=0.025 $Btu/(h \cdot ft^2 \cdot F)$
Windows	Double Glazing, Low e, U=0.416 $Btu/(h \cdot ft^2 \cdot F)$, SC=0.5
Windows to wall ratio (%)	19.2
Exterior and interior shades	None
Schedules	
Operation schedule	24/7
Secondary Systems	
Systems type	VAV with hot water reheat

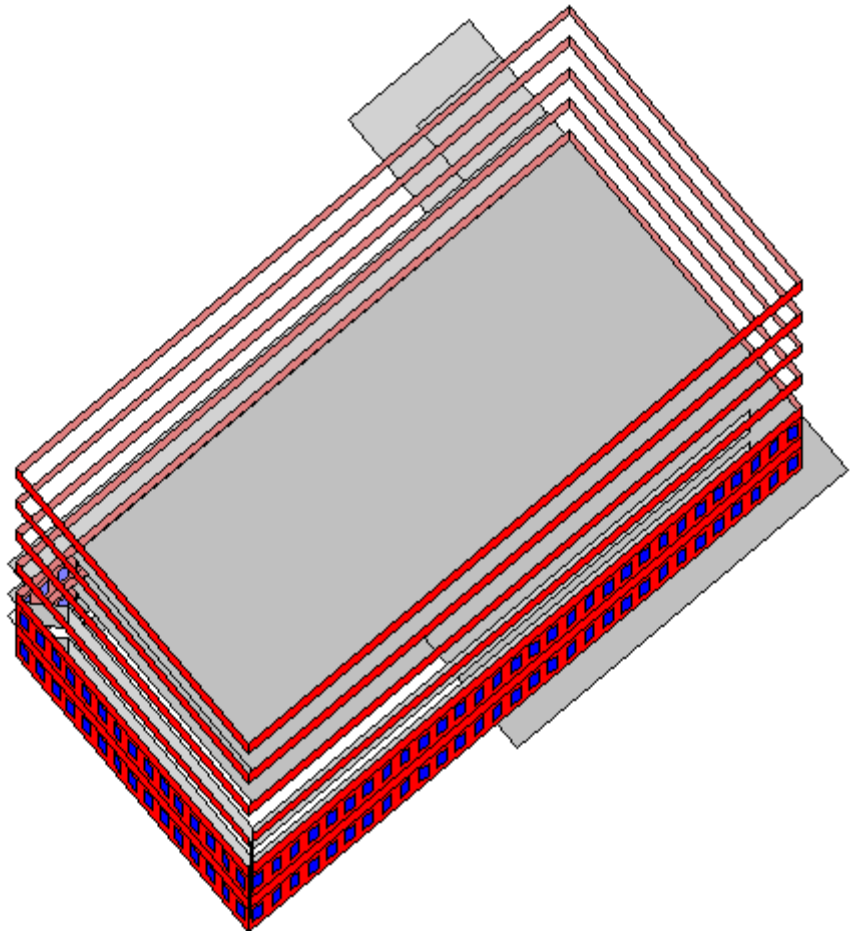


Figure A1. DOE2.1e DrawBDL for Large Hospital.

Table A2. Summary of Large Hotel Building Description

General	
Floor Area (ft^2)	619,200
Above Grade Floors	43
Below Grade Floors	0
% Conditioned and Lit	100
Geometry	
Footprint Shape	Square (120' X 120')
Zoning (1^{st} thru 4^{th} floor)	Each floor is a single zone
Zoning (5^{st} thru 42^{th} floor)	4 Perimeter/1 Interior
Zoning (43^{th} floor)	Single Zone
Perimeter Depth (feet)	20
Floor to floor height (ft)	13
Floor to ceiling height (ft)	9
Envelope	
Roof	Massive, R-27
Walls	Glass Curtain Wall, U=0.11 $Btu/(h \cdot ft^2 \cdot F)$
Foundation	Slab, U=0.025 $Btu/(h \cdot ft^2 \cdot F)$
Windows	Double Glazing, U=0.55 $Btu/(h \cdot ft^2 \cdot F)$, SC=0.41
Windows to wall ratio (%)	36
Exterior and interior shades	None
Schedules	
Operation schedule	24/7
Secondary Systems	
Lobby, conf. rooms, offices	VAV with hot water reheat
Guest rooms	Four pipe fans coils
Guest room – DOAS/Vent.	Reheat fan system
Mechanical room 43^{rd} floor	Single zone reheat

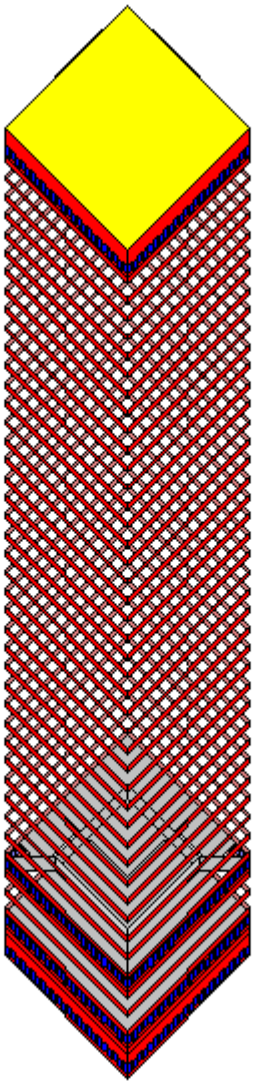


Figure A2. DOE2.1e DrawBDL for Large Hotel.

Table A3. Summary of Large Office Building Description

General	
Floor Area (ft^2)	588,000
Above Grade Floors	17
Below Grade Floors	0
% Conditioned and Lit	100
Buildings	
Typical office floor	340' X 100' (34,000 ft^2)
Mechanical penthouse	200' X 50' (10,000 ft^2)
Floor to floor height (ft)	13
Floor to ceiling height (ft)	9
Envelope	
Roof	Concrete 4 in 50% abs. 1 in. insulation
Walls	CMU grouted, 2 in., EIFS, 30% abs, U=0.1 $Btu/(h \cdot ft^2 \cdot F)$
Foundation	Slab, U=0.03 $Btu/(h \cdot ft^2 \cdot F)$
Windows	Double Glazing, Low e, U=0.416 $Btu/(h \cdot ft^2 \cdot F)$, SHGC=0.43
Windows to wall ratio (%)	29
Exterior and interior shades	None
Schedules	
Operation schedule	Per office schedules
Secondary Systems	
Office/Admin floors	VAV with hot water reheats
Mechanical room penthouse	Single zone reheat

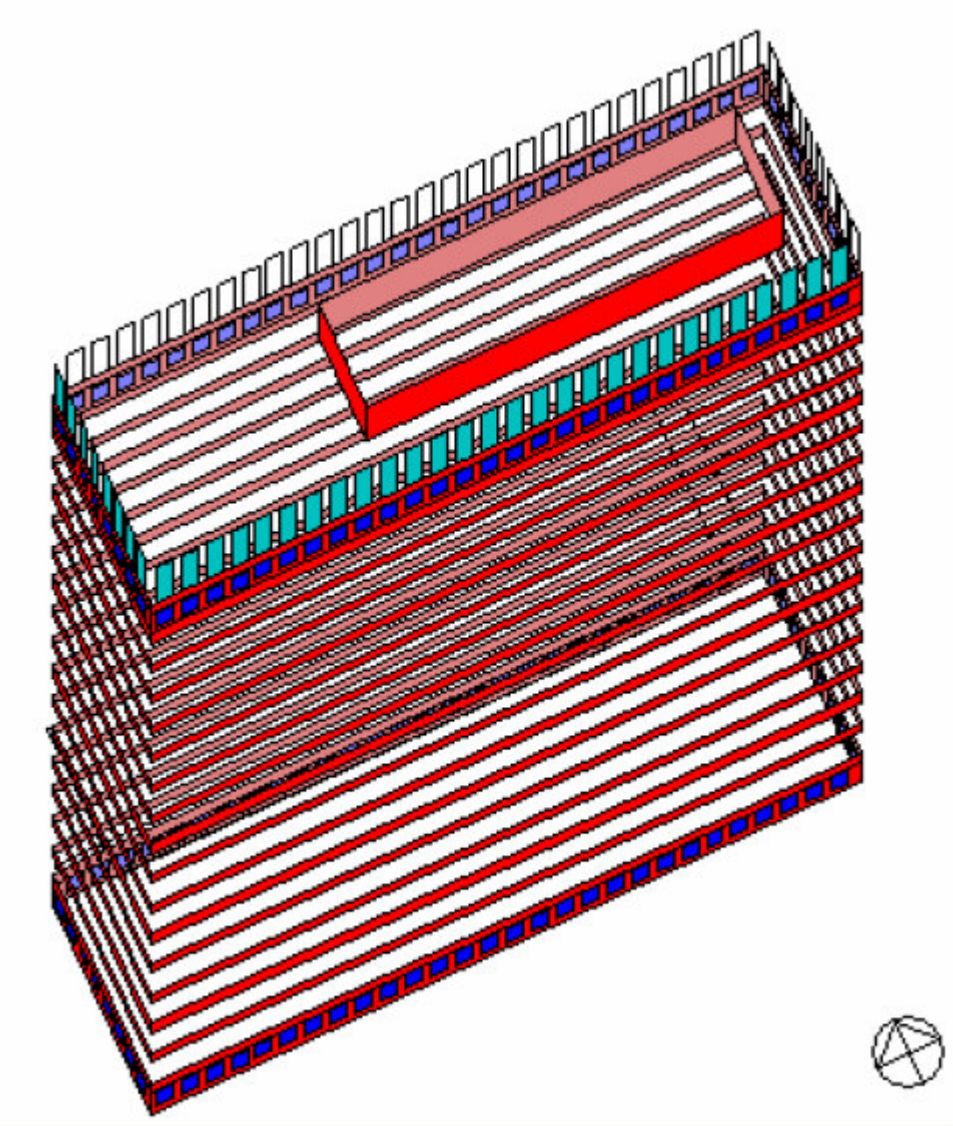


Figure A3. DOE2.1e DrawBDL for Large Office.

Table A4. Summary of Large School Building Description

General	
Floor Area (ft^2)	229,700
Above Grade Floors	Varies (3,2 and 1 depending on duty)
Below Grade Floors	0
% Conditioned and Lit	100
Buildings/Wings	
Classrooms	3 wings, 3 and 2 story (98,000 ft^2)
Auditorium	1 wing (12,600 ft^2)
Gymnasiums	2 wing (31,900 ft^2)
Cafeteria	1 wing (14,400 ft^2)
Office/Admin	1 annex (5,400 ft^2)
Central utility room	1 annex (5,400 ft^2)
Common (wings link)	62,000 ft^2
Floor to floor height (ft)	13 (typical), in Gymnasiums, Auditorium etc is higher
Floor to ceiling height (ft)	9 (typical)
Envelope	
Roof	Massive, R-25
Walls	CMU grouted, 2 in., EIFS, 30% abs, U=0.1 $Btu/(h \cdot ft^2 \cdot F)$
Foundation	Slab, U=0.03 $Btu/(h \cdot ft^2 \cdot F)$
Windows	Double Glazing, Low e, U=0.416 $Btu/(h \cdot ft^2 \cdot F)$,
Windows to wall ratio (%)	6.1
Exterior and interior shades	None
Schedules	
Operation schedule	Per office schedules
Secondary Systems	
Classrooms	Four Pipe Fan coils (FPFC)
Office/Admin, Common	VAV with hot water reheats
Auditorium, Gymnasiums,Cafeteria,	Single zone reheat

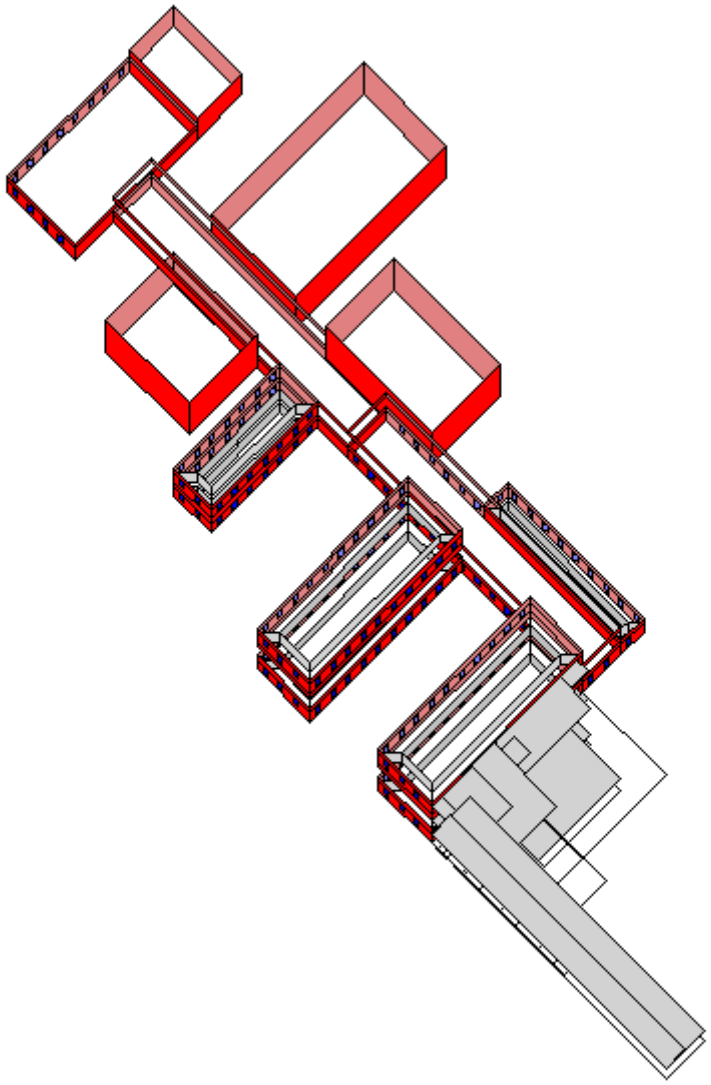


Figure A4. DOE2.1e DrawBDL for LargeSchool.

APPENDIX B: MATLAB 7 ROUTINE OF WAVELET ANALYSIS AND DAILY COOLING ENERGY USE MODELING FOR SHOOTOUT II COMPARISON

For the detailed algorithm and explanation of this routine, please refer to the technical reference manual:

ESL-ITR-08-11-02, Energy Systems Laboratory, Texas A&M University.

All the original building energy use data, weather data, routines and related Matlab toolbox are in the accompany CD-ROM of this manual.

```
nntwarn off;
clear all;

% Variable “neuron_number” stores number of neurons in hidden layer for each of the
3 % neural networks during each of 7 iterations, where each iteration represents an
input % case.

neuron_number=[ 1 1 2;
                1 1 2;
                1 1 2;
                1 1 1;
                1 1 2;
                1 1 2;
                2 1 1;];
nn_epochs=1000
goal=1e-3
dwtmode('per');

% Load Zachry building weather and energy use data file. Variable “A” contains all the
% days to be used for model training and testing which are determined by column 14 of
% variable “AA”. “n” is total number of days in “A” . Variable “AA” and “A” have the
% same format as the raw dataset.

[AA,B]=xlsread('Zachry_data.xls');
j=1;
```

```
for i=1:length(AA)
   if AA(i,14)==1
      A(j,:)=AA(i,:);
      j=j+1;
   end
end
n=length(A)/24;

% Generate variables for hourly OAT, RH, SOL and CWE. These data are in
% column 9, 10, 11 and 7 of variable "A" respectively.

for i=1:n
   oat(i,:)=A((i-1)*24+1:(i-1)*24+24,9)';
   rh(i,:)=A((i-1)*24+1:(i-1)*24+24,10)';
   sol(i,:)=A((i-1)*24+1:(i-1)*24+24,11)';
   cwe(i,:)=A((i-1)*24+1:(i-1)*24+24,7)';
end

% Apply cubic spline data interpolation to have the original 24 hour data interpolated
% to 32 data for discrete wavelet transforms.

x1=linspace(0,1,24);
x2=linspace(0,1,32);
for i=1:n
oat1(i,:)=spline(x1,oat(i,:),x2);
rh1(i,:)=spline(x1,rh(i,:),x2);
sol1(i,:)=spline(x1,sol(i,:),x2);
end

% Create variables for day of year and day types.

for i=1:n
   dayofyr(i)=A((i-1)*24+1,2);
   daytype(i)=A((i-1)*24+1,1);
end
daytype=daytype';

% Choose wavelet and decomposition level for OAT, RH and SOL.

wav_oat='db3'; level_oat=5;
wav_rh='db1';   level_rh=5;
wav_sol='db1'; level_sol=5;
```

```
% Apply DWT on OAT, RH and SOL. Store the resulting wavelet coefficients in
% variable “c_oat”, “c_rh” and “c_sol”.

for i=1:n
[c,Lo]=wavedec(oat1(i,:),level_oat,wav_oat);  c_oat(i,:)=c;
[c,Ld]=wavedec(rh1(i,:),level_rh,wav_rh);       c_rh(i,:)=c;
[c,Ls]=wavedec(sol1(i,:),level_sol,wav_sol);    c_sol(i,:)=c;
end


% Run 7 input cases to find the most significant coefficients for daily energy use
modeling. These coefficients will be used to define neighborhoods. “p” is neural
network model input for different cases.

for M2=1:7
   M2
   if M2==1
   p=c_oat(:,1);
   elseif M2==2
   p=[c_oat(:,1),c_rh(:,1)];
   elseif M2==3
   p=[c_oat(:,1),c_sol(:,1)];
   elseif M2==4
   p=[c_oat(:,1),c_rh(:,1),c_sol(:,1)];
   elseif M2==5
   p=[c_oat(:,1),c_oat(:,2)];
   elseif M2==6
   p=[c_oat(:,1),c_oat(:,2),c_rh(:,1)];
   elseif M2==7
   p=[c_oat(:,1),c_oat(:,2),c_sol(:,1)];
end

nnt1=neuron_number(M2,1);
nnt2=neuron_number(M2,2);
nnt3=neuron_number(M2,3);
% Determine the number of predictors for each case.

m=size(p);
m=m(2);

% Daily cooling energy use “t” is neural network model target (output).

t=sum(cwe,2);
```

```
% Classify the training and testing days into 3 groups of input/output pairs based on
% building operating schedules. Variable “p1” and “t1” are input and output data for
% ANN of group 1.

L1=0;L2=0;L3=0;
for i=1:n
   if daytype(i)==1
      L1=L1+1;
      p1(L1,:)=p(i,:);
      t1(L1)=t(i);
      dayofyr1(L1)=dayofyr(i);
   elseif daytype(i)==2
      L2=L2+1;
      p2(L2,:)=p(i,:);
      t2(L2)=t(i);
      dayofyr2(L2)=dayofyr(i);
   elseif daytype(i)==3
      L3=L3+1;
      p3(L3,:)=p(i,:);
      t3(L3)=t(i);
      dayofyr3(L3)=dayofyr(i);
   end
end

% Separate the days in each group into model training days and model testing days. 2/3
% of the days are used for training and 1/3 are used for testing. In this code, days are
% selected in time sequence, not randomly.

k1=ceil(L1*2/3);
for i=1:ceil(L1/3)
   temp1=(i-1)*3+1;
   temp2=(i-1)*3+2;
   temp3=(i-1)*3+3;
   day_trn((i-1)*2+1)=temp1;
   day_trn((i-1)*2+2)=temp2;
   day_tst(i)=temp3;
end

day_trn=day_trn(1:k1);
day_tst=day_tst(1:L1-k1);

% “p1_trn” and “t1_trn” are input and output data for ANN training for group 1.
% “p1_tst” and “t1_tst” are input and output data for ANN testing for group 1.
```

```
for i=1:k1
   p1_trn(i,:)=p1(day_trn(i),:);
   t1_trn(i)=t1(day_trn(i));
end
for i=1:L1-k1
   p1_tst(i,:)=p1(day_tst(i),:);
   t1_tst(i)=t1(day_tst(i));
end

k2=ceil(L2*2/3);
for i=1:ceil(L2/3)
   temp1=(i-1)*3+1;
   temp2=(i-1)*3+2;
   temp3=(i-1)*3+3;
   day_trn((i-1)*2+1)=temp1;
   day_trn((i-1)*2+2)=temp2;
   day_tst(i)=temp3;
end
day_trn=day_trn(1:k2);
day_tst=day_tst(1:L2-k2);

% “p2_trn” and “t2_trn” are input and output data for ANN training for group 2.
% “p2_tst” and “t2_tst” are input and output data for ANN testing for group 2.

for i=1:k2
   p2_trn(i,:)=p2(day_trn(i),:);
   t2_trn(i)=t2(day_trn(i));
end
for i=1:L2-k2
   p2_tst(i,:)=p2(day_tst(i),:);
   t2_tst(i)=t2(day_tst(i));
end

k3=ceil(L3*2/3);
for i=1:ceil(L3/3)
   temp1=(i-1)*3+1;
   temp2=(i-1)*3+2;
   temp3=(i-1)*3+3;
   day_trn((i-1)*2+1)=temp1;
   day_trn((i-1)*2+2)=temp2;
   day_tst(i)=temp3;
end
day_trn=day_trn(1:k3);
day_tst=day_tst(1:L3-k3);
```

```
% “p3_trn” and “t3_trn” are input and output data for ANN training for group 3.
% “p3_tst” and “t3_tst” are input and output data for ANN testing for group 3.

for i=1:k3
   p3_trn(i,:)=p3(day_trn(i),:);
   t3_trn(i)=t3(day_trn(i));
end
for i=1:L3-k3
   p3_tst(i,:)=p3(day_tst(i),:);
   t3_tst(i)=t3(day_tst(i));
end

p1=p1'; p2=p2'; p3=p3';
p1_trn=p1_trn'; p2_trn=p2_trn'; p3_trn=p3_trn';
p1_tst=p1_tst'; p2_tst=p2_tst'; p3_tst=p3_tst';

% Create a feed-forward back-propagation neural network for group 1.

[pn1_trn,meanp1,stdp1,tn1_trn,meant1,stdt1]=prestd(p1_trn,t1_trn);
net=newff(minmax(pn1_trn),[nnt1,1],{'tansig','purelin'},'trainlm','learngdm','mse');
net.trainParam.show=100;
net.trainParam.epochs=nn_epochs;
net.trainParam.goal=goal;

% Train the neural network “net1” of group 1.

net1=train(net,pn1_trn,tn1_trn);

% Calculate CV and MBE of energy modeling for training dataset in group 1.

cwe_pre1_trn=sim(net1,pn1_trn);
cwe_pre1_trn=poststd(cwe_pre1_trn,meant1,stdt1);
cwe_ori1_trn=t1_trn;

error1=cwe_pre1_trn-cwe_ori1_trn;
cv1_trn=(sum(error1.^2)/k1)^0.5/mean(cwe_ori1_trn)
mbe1_trn=(sum(error1)/k1)/mean(cwe_ori1_trn);

% Calculate CV and MBE of energy modeling for testing dataset in group 1.

[pn1_tst]=trastd(p1_tst,meanp1,stdp1);
[tn1_tst]=trastd(t1_tst,meant1,stdt1);
cwe_pre1_tst=sim(net1,pn1_tst);
cwe_pre1_tst=poststd(cwe_pre1_tst,meant1,stdt1);
```

```
cwe_ori1_tst=t1_tst;

error1_tst=cwe_pre1_tst-cwe_ori1_tst;
cv1_tst=(sum(error1_tst.^2)/(L1-k1-m))^0.5/mean(cwe_ori1_tst)
mbe1_tst=(sum(error1_tst)/(L1-k1-m))/mean(cwe_ori1_tst);

% Modeled energy use for all the days in group 1.

[pn1]=trastd(p1,meanp1,stdp1);
cwe_pre1=sim(net1,pn1);
cwe_pre1=poststd(cwe_pre1,meant1,stdt1);

% Create a feed-forward back-propagation network for group 2.

[pn2_trn,meanp2,stdp2,tn2_trn,meant2,stdt2]=prestd(p2_trn,t2_trn);
net=newff(minmax(pn2_trn),[nnt2,1],{'tansig','purelin'},'trainlm','learngdm','mse');
net.trainParam.show=100;
net.trainParam.epochs=nn_epochs;
net.trainParam.goal=goal;

% Train the neural network “net2” of group 2.

net2=train(net,pn2_trn,tn2_trn);

% Calculate CV and MBE of energy modeling for training dataset in group 2.

cwe_pre2_trn=sim(net2,pn2_trn);
cwe_pre2_trn=poststd(cwe_pre2_trn,meant2,stdt2);
cwe_ori2_trn=t2_trn;

error2=cwe_pre2_trn-cwe_ori2_trn;
cv2_trn=(sum(error2.^2)/k2)^0.5/mean(cwe_ori2_trn)
mbe2_trn=(sum(error2)/k2)/mean(cwe_ori2_trn);

% Calculate CV and MBE of energy modeling for testing dataset in group 2.

[pn2_tst]=trastd(p2_tst,meanp2,stdp2);
[tn2_tst]=trastd(t2_tst,meant2,stdt2);
cwe_pre2_tst=sim(net2,pn2_tst);
cwe_pre2_tst=poststd(cwe_pre2_tst,meant2,stdt2);
cwe_ori2_tst=t2_tst;

error2_tst=cwe_pre2_tst-cwe_ori2_tst; % t is measured cwe
cv2_tst=(sum(error2_tst.^2)/(L2-k2-m))^0.5/mean(cwe_ori2_tst)
```

```
mbe2_tst=(sum(error2_tst)/(L2-k2-m))/mean(cwe_ori2_tst);

% Modeled energy use for all the days in group 2.

 [pn2]=trastd(p2,meanp2,stdp2);
cwe_pre2=sim(net2,pn2);
cwe_pre2=poststd(cwe_pre2,meant2,stdt2);

% Create a feed-forward back-propagation network for group 3.

[pn3_trn,meanp3,stdp3,tn3_trn,meant3,stdt3]=prestd(p3_trn,t3_trn);
net=newff(minmax(pn3_trn),[nnt3,1],{'tansig','purelin'},'trainlm','learngdm','mse');
net.trainParam.show=100;
net.trainParam.epochs=nn_epochs;
net.trainParam.goal=goal;

% Train neural network “net3” of group 3.

net3=train(net,pn3_trn,tn3_trn);

% Calculate CV and MBE of energy modeling for training dataset in group 3.

cwe_pre3_trn=sim(net3,pn3_trn);
cwe_pre3_trn=poststd(cwe_pre3_trn,meant3,stdt3);
cwe_ori3_trn=t3_trn;

error3=cwe_pre3_trn-cwe_ori3_trn;
cv3_trn=(sum(error3.^2)/k3)^0.5/mean(cwe_ori3_trn)
mbe3_trn=(sum(error3)/k3)/mean(cwe_ori3_trn);


% Calculate CV and MBE of energy modeling for testing dataset in group 3.

[pn3_tst]=trastd(p3_tst,meanp3,stdp3);
[tn3_tst]=trastd(t3_tst,meant3,stdt3);
cwe_pre3_tst=sim(net3,pn3_tst);
cwe_pre3_tst=poststd(cwe_pre3_tst,meant3,stdt3);
cwe_ori3_tst=t3_tst;

error3_tst=cwe_pre3_tst-cwe_ori3_tst; % t is measured cwe
cv3_tst=(sum(error3_tst.^2)/(L3-k3-m))^0.5/mean(cwe_ori3_tst)
mbe3_tst=(sum(error3_tst)/(L3-k3-m))/mean(cwe_ori3_tst);

% Modeled energy use for all the days in group 3.
```

```
 [pn3]=trastd(p3,meanp3,stdp3);
cwe_pre3=sim(net3,pn3);
cwe_pre3=poststd(cwe_pre3,meant3,stdt3);

% Calculate CV of energy modeling for all the training days and all the testing days
% respectively.

cweoftrn=[t1_trn' cwe_pre1_trn';t2_trn' cwe_pre2_trn';t3_trn' cwe_pre3_trn'];
cweoftst=[t1_tst' cwe_pre1_tst';t2_tst' cwe_pre2_tst';t3_tst' cwe_pre3_tst'];
cv_trn=(sum((cweoftrn(:,1)-cweoftrn(:,2)).^2)/(k1+k2+k3-m))^0.5/mean(cweoftrn(:,1))
cv_tst=(sum((cweoftst(:,1)-cweoftst(:,2)).^2)/(n-k1-k2-k3-m))^0.5/mean(cweoftst(:,1))

% Calculate CV of energy modeling for the whole dataset.

temp=[dayofyr1' t1' cwe_pre1' (cwe_pre1-t1)';dayofyr2' t2' cwe_pre2' (cwe_pre2-
t2)';dayofyr3' t3' cwe_pre3' (cwe_pre3-t3)'];
temp=sortrows(temp,1);
cv_total=(sum((temp(:,2)-temp(:,3)).^2)/(n-m))^0.5/mean(temp(:,2))

% Generate output variable “CV” which contains all the CVs for 7 cases.

CV(M2,1)=cv_trn;
CV(M2,2)=cv_tst;
CV(M2,3)=cv_total;
CV(M2,4)=(cv_trn-cv_tst)/cv_trn;

clear p1 p1_trn p1_tst t1 t1_trn t1_tst
clear p2 p2_trn p2_tst t2 t2_trn t2_tst
clear p3 p3_trn p3_tst t3 t3_trn t3_tst
clear dayofyr1 dayofyr2 dayofyr3 dayofyr1_trn dayofyr1_tst
end
```

APPENDIX C: MATLAB 7 ROUTINE OF NEIGHBORHOOD CLASSIFICATION FOR COOLING ENERGY USE MODELING FOR SHOOTOUT II COMPARISON

For the detailed algorithm and explanation of this routine, please refer to the technical reference manual:

ESL-ITR-08-11-02, Energy Systems Laboratory, Texas A&M University.

All the original building energy use data, weather data, routines and related Matlab toolbox are in the accompany CD-ROM of this manual.

```
Clear all;
nntwarn off;
dwtmode('per');

% Load Zachry building weather and energy use data file.
% "n" is total number of days in "A".

[A,B]=xlsread('Zachry_data.xls');
n=length(A)/24;

% Generate variables for hourly OAT, RH, SOL and CWE. These data are in
% column 9, 10, 11 and 7 of variable "A" respectively.

For i=1:n
   oat(i,:) =A((i-1)*24+1:(i-1)*24+24,9)';
   rh(i, :) =A((i-1)*24+1:(i-1)*24+24,10)';
   sol(i, :) =A((i-1)*24+1:(i-1)*24+24,11)';
   cwe(i, :) =A((i-1)*24+1:(i-1)*24+24,7)';
end

% Calculate daily cooling energy use
cwe_daily=sum(cwe,2);
```

```
% Apply cubic spline data interpolation to have the original 24 hour data interpolated
% to 32 data for discrete wavelet transforms.
X1=linspace(0,1,24);
x2=linspace(0,1,32);
for i=1:n
oat1(i,:)=spline(x1,oat(i,:),x2);
rh1(i,:)=spline(x1,rh(i,:),x2);
sol1(i,:)=spline(x1,sol(i,:),x2);
end

% Create variable for day of year.

For i=1:n
   dayofyr(i)=A((i-1)*24+1,2);
end

% Identify day of year for days with complete information which are used for model
% training and testing. This is determined by column 14 of variable "A".

k1=0;
for i=1:n
   j=i*24-23;
   if A(j,14)==1
      k1=k1+1;
      dayofyr_trntst(k1)=A(j,2);
   end
end

% Choose wavelet and decomposition level for OAT, RH and SOL.

Wav_oat='db3'; step_oat=5;
wav_rh='db1';  step_rh=5;
wav_sol='db1'; step_sol=5;

% Apply DWT on OAT, RH and SOL. Store the resulting wavelet coefficients in
% variable "c_oat", "c_rh" and "c_sol".

For i=1:n
[c,Lo]=wavedec(oat1(I, i,:),step_oat,wav_oat);    c_oat(I, i,:)=c;
[c,Lh]=wavedec(rh1(I, i,:),step_rh,wav_rh);         c_rh(I, i,:)=c;
[c,Ls]=wavedec(sol1(I, i,:),step_sol,wav_sol);      c_sol(I, i,:)=c;
end

% Calculate weights of the significant wavelet coefficients through multiple linear
```

```
% regression of the significant wavelet coefficients against CWE.

Coef_cwe=[c_oat(:,1),c_oat(:,2),c_rh(:,1)];
for i=1:k1
   temp1(i)=cwe_daily(dayofyr_trntst(i));
   temp2(i,:)=coef_cwe(dayofyr_trntst(i),:);
end
x=[ones(k1,1) temp2(:,1) temp2(:,2) temp2(:,3)];
y=temp1';
a=x\y;
coeff_weight=a(2:end);
coeff_weight=abs(coeff_weight)



% Apply weights to the corresponding significant wavelet coefficients.
% Generate the variable "SOM_data" as Self-Organizing Map input data.

SOM_data=coef_cwe;
temp3=size(SOM_data);
temp3=temp3(2);
for i=1:temp3
   SOM_data(:,i)=SOM_data(:,i)*coeff_weight(i);
end

% Create and format data structure from "SOM_data".

Sd=som_data_struct(SOM_data);

% Create, initialize and train the SOM

sm=som_make(sd,'msize',[15 15]);

% U-matrix (unified distance matrix) visualization of the trained SOM.

Som_show(sm,'umat','all');hold on
som_grid(sm);hold off
```

APPENDIX D: MATLAB 7 ROUTINE OF HOURLY COOLING ENERGY USE PREDICTION MODEL FOR SHOOTOUT II COMPARISON

For the detailed algorithm and explanation of this routine, please refer to the technical reference manual:

ESL-ITR-08-11-02, Energy Systems Laboratory, Texas A&M University.

All the original building energy use data, weather data, routines and related Matlab toolbox are in the accompany CD-ROM of this manual.

```
Clear all;
nntwarn off;

nn_epochs=1000
som_epochs=100
goal=1e-3
dwtmode('per');
state1=2;

% Number of neighborhoods.

node1=1;
node2=3;
n_nb=node1*node2

%  Variable “nntnb” contains number of hidden neurons of ANN for three day types.

nntnb=[1 4 1]

% Load Zachry building weather and energy use data file.
% “n” is total number of days in “A”.

[AA,B]=xlsread('Zachry_data.xls');
A=AA;
n=length(A)/24;
```

```
% Generate variables for hourly OAT, RH, SOL and CWE. These data are in
% column 9, 10, 11 and 7 of variable "A" respectively.

for i=1:n
   oat(i,:)=A((i-1)*24+1:(i-1)*24+24,9)';
   rh(i,:)=A((i-1)*24+1:(i-1)*24+24,10)';
   sol(i,:)=A((i-1)*24+1:(i-1)*24+24,11)';
   cwe(i,:)=A((i-1)*24+1:(i-1)*24+24,7)';
end

% Apply cubic spline data interpolation to have the original 24 hour data interpolated
% to 32 data for discrete wavelet transforms.

x1=linspace(0,1,24);
x2=linspace(0,1,32);
for i=1:n
oat1(i,:)=spline(x1,oat(i,:),x2);
rh1(i,:)=spline(x1,rh(i,:),x2);
sol1(i,:)=spline(x1,sol(i,:),x2);
end

% Calculate daily average OAT, RH, SOL and CWE.

oat_avg=sum(oat1,2)/32;
rh_avg=sum(rh1,2)/32;
sol_avg=sum(sol1,2)/32;
cwe_daily=sum(cwe,2);

% Create variables for day of year and day types.

for i=1:n
   dayofyr(i)=A((i-1)*24+1,2);
   daytype(i)=A((i-1)*24+1,1);
end
daytype=daytype';

% Identify day of year for days with complete information which are used for model
% training and testing and day of year for days with incomplete energy use data which
% are to be predicted by the trained model. This is determined by column 14 of variable
% "A".

k1=0;k2=0;
for i=1:n
   j=i*24-23;
```

```
   if A(j,14)==1
      k1=k1+1;
      dayofyr_trntst(k1)=A(j,2);
   else
      k2=k2+1;
      dayofyr_pre(k2)=A(j,2);
   end
end

% Choose wavelet and decomposition level for OAT, RH and SOL.

wav_oat='db3'; step_oat=5;
wav_rh='db1';  step_rh=5;
wav_sol='db1'; step_sol=5;

% Apply DWT on OAT, RH and SOL. Store the resulting wavelet coefficients in
% variable “c_oat”, “c_rh” and “c_sol”.

for i=1:n
[c,Lo]=wavedec(oat1(i,:),step_oat,wav_oat);  c_oat(i,:)=c;
[c,Lh]=wavedec(rh1(i,:),step_rh,wav_rh);       c_rh(i,:)=c;
[c,Ls]=wavedec(sol1(i,:),step_sol,wav_sol);    c_sol(i,:)=c;
end

% Calculate weights of the significant wavelet coefficients through multiple linear
% regression of the significant wavelet coefficients against CWE.

for i=1:k1
   temp1(i)=cwe_daily(dayofyr_trntst(i));
   temp2(i,:)=coef_cwe(dayofyr_trntst(i),:);
end
x=[ones(k1,1) temp2(:,1) temp2(:,2) temp2(:,3)];
y=temp1';
a=x\y;
coeff_weight=a(2:end);
coeff_weight=abs(coeff_weight)

% Apply weights to the corresponding significant wavelet coefficients.
% Generate the variable “p” as Self-Organizing Map input data.

p=coef_cwe;
num4=size(p);num4=num4(2);
for i=1:num4
   p(:,i)=p(:,i)*coeff_weight(i);
```

```
end

% Train the SOM with input “p”.

p=p';
net=newsom(minmax(p),[node1 node2]);
net.trainParam.epochs=som_epochs;
net=train(net,p);

% Identify the neighborhood of each day it belongs to. Variable “b” contains this
% information. Variable “nb_doy” contains day of year in each neighborhood.

for i=2:n
    a=sim(net,p(:,i));
    m=find(a==1);
    b(i,1:2)=[dayofyr(i) m];
end

for i=1:n_nb
   nb{i}=find(b(:,2)==i)';
   s2=length(nb{i});
   for j=1:s2
      nb_doy{i}(j)=b(nb{i}(j),1);
   end
end

clear p;
clear temp;

for i=1:n_nb
    temp(i,1)=i;
    temp(i,2)=length(nb{i});
end
temp=sortrows(temp,2);
for i=1:n_nb
    nb_new{i}=nb{temp(i)};
    nb_doy_new{i}=nb_doy{temp(i)};
end
for i=1:n_nb
    nb{i}=nb_new{i};
    nb_doy{i}=nb_doy_new{i};
end
for i=2:n
    x=find(temp(:,1)==b(i,2));
```

```
   b(i,2)=x;
end
% Determine days for model training, testing and energy prediction in each neighborhood.
% Variable "nb_doy_trntst" contains days for training and testing in each neighborhood.
% Variable "nb_doy_pre" contains days for energy use prediction in each neighborhood.

for i=1:n_nb
   k3=0;k4=0;
   for j=1:length(nb_doy{i})
      temp=nb_doy{i}(j);
      if A(temp*24-23,14)==1
         k3=k3+1;
         nb_doy_trntst{i}(k3)=temp;
      elseif A(temp*24-23,14)==0
         k4=k4+1;
         nb_doy_pre{i}(k4)=temp;
      end
   end
end

p_trn=cell(1,n_nb);t_trn=cell(1,n_nb);p_tst=cell(1,n_nb);t_tst=cell(1,n_nb);
meanp=cell(1,n_nb);meant=cell(1,n_nb);stdp=cell(1,n_nb);stdt=cell(1,n_nb);
minp=cell(1,n_nb);maxp=cell(1,n_nb);mint=cell(1,n_nb);maxt=cell(1,n_nb);
cwe_pre_trn=cell(1,n_nb);cwe_pre_tst=cell(1,n_nb);

% Create input/output pairs for model training for each neighborhood.
% Use 2/3 of the training and testing days for training.
% Variable "p_trn" is model input and "t_trn" is model output.

for i=1:n_nb
   L(i)=length(nb_doy_trntst{i});
   k(i)=ceil(L(i)*2/3);
   day=randdeintrlv(1:L(i),state1);
   temp1=sort(day(1:k(i)),2);                    % Training dataset, 2/3 of total
   temp2=sort(day(k(i)+1:end),2);                % Testing dataset, 1/3 of total
 for j=1:k(i)
   k1=nb_doy_trntst{i}(temp1(j));
   p_trn{i}(1,(j-1)*24+1:j*24)=AA((k1-1)*24:(k1-1)*24+23,9)';   % previous hour OAT
   p_trn{i}(2,(j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,9)';   % current OAT
   p_trn{i}(3,(j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,10)';   % current RH
   p_trn{i}(4,(j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,11)';   % current SOL
   p_trn{i}(5,(j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,6)';     % current HOUR
   p_trn{i}(6,(j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,1)';     % DAY TYPE
   p_trn{i}(7,(j-1)*24+1:j*24)=AA((k1-1)*24:(k1-1)*24+23,7)';  % previous hour CWE
```

```
   t_trn{i}((j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,7)';     % current hour CWE
  end
% If the previous hour CWE is unknown (which is 1000000), just substitute by the
% current hour CWE.

  for j=1:k(i)*24
      if p_trn{i}(7,j)==1000000
         p_trn{i}(7,j)=t_trn{i}(j);
      end
  end

% Create input/output pairs for model testing for each neighborhood.
% Use 1/3 of the training and testing days for testing.
% Variable "p_tst" is model input and "t_tst" is model output.

  for j=1:L(i)-k(i)
    k1=nb_doy_trntst{i}(temp2(j));
    p_tst{i}(1,(j-1)*24+1:j*24)=AA((k1-1)*24:(k1-1)*24+23,9)';    % previous hour OAT
    p_tst{i}(2,(j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,9)';   % current OAT
    p_tst{i}(3,(j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,10)';  % current RH
    p_tst{i}(4,(j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,11)';  % current SOL
    p_tst{i}(5,(j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,6)';    % current HOUR
    p_tst{i}(6,(j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,1)';    % DAY TYPE
    p_tst{i}(7,(j-1)*24+1:j*24)=AA((k1-1)*24:(k1-1)*24+23,7)';   % previous hour CWE
    t_tst{i}((j-1)*24+1:j*24)=AA((k1-1)*24+1:(k1-1)*24+24,7)';   % current hour CWE
    end

% If the previous hour CWE is unknown (which is 1000000), just substitute by the
% current hour CWE.

  for j=1:(L(i)-k(i))*24
      if p_tst{i}(7,j)==1000000
         p_tst{i}(7,j)=t_tst{i}(j);
      end
  end
end

% Number of predictors.
w=7;

% Create and train ANN model for each neighborhood

nnet=cell(1,n_nb);
for i=1:n_nb
```

```
i
pp=p_trn{i};
tt=t_trn{i};
[pn,meanp{i},stdp{i},tn,meant{i},stdt{i}] = prestd(pp,tt);
net=newff(minmax(pn),[nntnb(i),1],{'tansig','purelin'},'trainlm','learngdm','mse');

% Initialize ANN.

 [p1,q1]=size(net.IW{1,1});
net.IW{1,1}=0.5*ones(p1,q1);
[p2,q2]=size(net.LW{2,1});
net.LW{2,1}=0.5*ones(p2,q2);
[p3,q3]=size(net.b{1});
net.b{1}=ones(p3,q3);
[p4,q4]=size(net.b{2});
net.b{2}=ones(p4,q4);

net.trainParam.show=100;
net.trainParam.epochs=nn_epochs;
net.trainParam.goal=goal;
net=train(net,pn,tn);
weight=net.iw{1};
nnet{i}=net;
clear pp tt;

% Simulated energy use for training days and testing days.
% Variable “cwe_pre_trn” and  “cwe_pre_tst”  are model simulated energy use for
% training days and testing days respectively in each neighborhood.

pp=p_trn{i};
[pn]=trastd(pp,meanp{i},stdp{i});
temp=sim(nnet{i},pn);
cwe_pre_trn{i}=poststd(temp,meant{i},stdt{i});

pp=p_tst{i};
[pn]=trastd(pp,meanp{i},stdp{i});
temp=sim(nnet{i},pn);
cwe_pre_tst{i}=poststd(temp,meant{i},stdt{i});

% Calculate CV and MBE of energy modeling for all the training days and all the testing
% days respectively in each neighborhood.

error=cwe_pre_trn{i}-t_trn{i};
cv_trn=(sum(error.^2)/(k(i)*24-w))^0.5/mean(t_trn{i})
```

```
mbe_trn=(sum(error)/(k(i)*24-w))/mean(t_trn{i});

error=cwe_pre_tst{i}-t_tst{i};
cv_tst=(sum(error.^2)/((L(i)-k(i))*24-w))^0.5/mean(t_tst{i})
mbe_tst=(sum(error)/((L(i)-k(i))*24-w))/mean(t_tst{i});

end

% Calculate CV and MBE of energy modeling for all the training days, all the testing
days and combined training and testing days in the raw dataset.
% “cv_trn_total” and “mbe_trn_total” are CV and MBE for energy modeling of all the
% training days.
% “cv_tst_total” and “mbe_tst_total” are CV and MBE for energy modeling of all the
% testing days.
% “cv_total” and “mbe_total” are CV and MBE for energy modeling of all the combined
% trainging and testing days.

temp1=1;temp2=0;
for i=1:n_nb
   temp2=temp2+length(t_trn{i});
   trn(temp1:temp2,1)=t_trn{i}';
   trn(temp1:temp2,2)=cwe_pre_trn{i}';
   trn(temp1:temp2,3)=cwe_pre_trn{i}'-t_trn{i}';
   temp1=temp2+1;
end
cv_trn_total=(sum(trn(:,3).^2)/(length(trn)-w))^0.5/mean(trn(:,1))
mbe_trn_total=(sum(trn(:,3))/(length(trn)-w))/mean(trn(:,1));

temp1=1;temp2=0;
for i=1:n_nb
   temp2=temp2+length(t_tst{i});
   tst(temp1:temp2,1)=t_tst{i}';
   tst(temp1:temp2,2)=cwe_pre_tst{i}';
   tst(temp1:temp2,3)=cwe_pre_tst{i}'-t_tst{i}';
   temp1=temp2+1;
end
cv_tst_total=(sum(tst(:,3).^2)/(length(tst)-w))^0.5/mean(tst(:,1))
mbe_tst_total=(sum(tst(:,3))/(length(tst)-w))/mean(tst(:,1));

trntst=[trn;tst];
cv_total=(sum(trntst(:,3).^2)/(length(trntst)-w))^0.5/mean(trntst(:,1))
mbe_total=(sum(trntst(:,3))/(length(trntst)-w))/mean(trntst(:,1));
```

```
% Energy use prediction.

% Identify hour of year with unknown energy use (which is 1000000 in column 7 of raw
% dataset).
% “j” is day of year for the hour with unknown energy use.
% “s1” is the neighborhood the day belongs to.

clear pp
L1=0;
for i=25:n*24
   if A(i,7)==1000000
      L1=L1+1;
      j=1+floor((i-1)/24);
      s1=b(j,2);
meanp1=meanp{s1};meant1=meant{s1};stdp1=stdp{s1};stdt1=stdt{s1};

pp(1)=A(i-1,9);    % previous hour OAT
pp(2)=A(i,9);       % OAT
pp(3)=A(i,10);     % RH
pp(4)=A(i,11);     % SOL
pp(5)=A(i,6);       % HOUR
pp(6)=A(i,1);       % DAY TYPE
pp(7)=A(i-1,7);    % previous hour CWE
pp=pp';
[pn]=trastd(pp,meanp1,stdp1);
cwe_pre(L1)=sim(nnet{s1},pn);
cwe_pre(L1)=poststd(cwe_pre(L1),meant1,stdt1);
cwe_ans(L1)=AA(i,12);
clear pp;

% Fill the missing energy use with the predicted energy use in dataset “A”.
% The predicted energy use will be used as time-lagged data for next hour energy use
% prediction.

A(i,7)=cwe_pre(L1);
   end
end

% Remove the data points that measured energy uses are not given.

M=0;
for i=1:L1
   if cwe_ans(i)~=1000000
      M=M+1;
```

```
        cwe_pre1(M)=cwe_pre(i);
        cwe_ans1(M)=cwe_ans(i);
    end
end

% Calculate CV and MBE of energy prediction for unknown energy use.

error=cwe_pre1-cwe_ans1;
cv=(sum(error.^2)/(M-w))^0.5/mean(cwe_ans1)
mbe=(sum(error)/(M-w))/mean(cwe_ans1)
```

APPENDIX E: MATLAB 7 ROUTINE OF CHANGE-POINT MODEL

```
%load cooling and heating training data and testing data
load cp_data;
data=cp_cooling_trn; %cooling energy model training data
n=length(data);

for i=3:n-3
   data1=data(1:i,:);data2=data(i+1:end,:);
   b1(i)=(dot(data1(:,1),data1(:,2))-
(sum(data1(:,1))*sum(data1(:,2)))/i)/(sum(data1(:,1).^2)-sum(data1(:,1))^2/i);
   a1(i)=mean(data1(:,2))-b1(i)*mean(data1(:,1));

   %Find change point x0

   hf=(sum(data2(:,1)))^2-(n-i)*sum(data2(:,1).^2);
   x0(i)=(sum(data2(:,2))*sum(data2(:,1).^2)-
dot(data2(:,1),data2(:,2))*sum(data2(:,1))+a1(i)*hf)/(sum(data2(:,1))*sum(data2(:,2))-
(n-i)*dot(data2(:,1),data2(:,2))-b1(i)*hf);
   y0(i)=a1(i)+b1(i)*x0(i);
   b2(i)=(sum(data2(:,2))-(n-i)*y0(i))/(sum(data2(:,1))-(n-i)*x0(i));
   a2(i)=y0(i)-b2(i)*x0(i);

   %calculae residual sum of squares RSS

   RSS(i)=sum((data1(:,2)-(a1(i)+b1(i)*data1(:,1))).^2)+sum((data2(:,2)-
(a2(i)+b2(i)*data2(:,1))).^2);
   residual(:,i)=[data1(:,2)-(a1(i)+b1(i)*data1(:,1));data2(:,2)-(a2(i)+b2(i)*data2(:,1))];

   if x0(i)<data1(end,1) | x0(i)>data2(1,1)
      RSS(i)=100000000;
   end
end

p=find(RSS==min(RSS(3:end))) %change point p
a1=a1(p)
b1=b1(p)
a2=a2(p)
b2=b2(p)
%calculate CV and MBE :
```

```
cv=(RSS(p)/n)^0.5/mean(data(:,2))
mbe=(sum(residual(p))/n)/mean(data(:,2))
```

VITA

Yafeng Lei was born in Hunan Province, China. He received his Bachelor of Engineering degree in thermal energy engineering in 1997 from Tianjin University and then worked in industry as an engineer for five years. He enrolled in Texas A&M University in August 2002. After receiving his Master of Science degree in 2005, he continued his Ph.D. study under the guidance of Dr. Kris Subbarao at Texas A&M University. He received his Ph.D. in August 2009.

Yafeng Lei can be reached through Department of Mechanical Engineering, 3123 TAMU, College Station, TX 77843-3123. His email address is: leiyafeng@hotmail.com.