Data Processing and Data Mining on Energy Consumption Database of Commercial Buildings in Shanghai

Yiqun Pan, PhD
Member ASHRAE

Zhizhong Huang

Xiaowei Zheng

ABSTRACT

This paper adopts data processing methods and data mining technology to develop a building energy consumption model, based on an energy consumption database of commercial buildings that includes 95 commercial buildings in Shanghai. Data transformation and data reduction are conducted to clear up data relations in the database. Three methods for missing data handling as well as outlier inspection are used for data processing. The software SAS is used as the tool for data processing and data mining. An optimum regression model of building energy consumption is made for each missing data element. Through comparing the three optimum regression models and their prediction results of building energy consumption, it is found that the Regression Imputation Method was the best method to handle missing data, and a regression model with operation time of HVAC system, cooling capacity, ratio of office area to total gross area, and hotel area to total gross area was the most reasonable prediction model of the energy consumption of commercial buildings in Shanghai.

INTRODUCTION

In recent years, the energy consumed by buildings has been increasing rapidly in China both in the absolute quantity and the ratio in national total energy consumption, and it is certainly becoming a main field of energy saving in China.

A lot of work has been conducted by the professionals around the world on investigation, statistics and analysis on building energy consumption and the programming of building energy simulation software. Many researchers are dedicated in the statistics, analysis and modeling of building energy consumption using regression methodology. Lam et al. (1997) used the simulation computer program DOE-2 (LBNL 1991) to carry out a parametric study of a typical high-rise air-conditioned office building in Hong Kong. A total of 28 relative design parameters were found to correlate well with the predicted annual electricity consumption. Both linear and non-linear multiple regression techniques were used to develop regression models and energy equations for the prediction of annual electricity use. Twelve input design parameters were found to be the most significant design variables and were used in the energy prediction equations. General regression neural networks (GRNN) were adopted by Ben-Nakhi and Mahmoud (2004) to optimize HVAC thermal energy storage in public buildings as well as office buildings. The training database for the GRNN was generated using the building simulation software ESP-r. Three different buildings were investigated, and hourly outdoor temperatures and building cooling loads were the input and the output of the GRNN, respectively. The results showed that a properly designed NN is a powerful instrument for optimizing thermal energy storage in buildings, and it can work well when based only on outdoor temperature records. Chung et al. (2006) described a benchmarking process for energy efficiency by means of multiple regression analysis, with which the relationship between energy-use intensities (EUIs) and the explanatory factors (e.g., operating hours) is developed. Nine variables including building age, occupancy and type of energy system are adopted to establish the regression equation, and a benchmarking table is derived by removing the effect of variance in the significant explanatory factors using the multiple-regression model. Ghiaus (2006) adopted energy consumption and outdoor temperature recorded by a Building Energy Management Systems (BEMS) to assess the energy performance of

Yiqun Pan is a professor and **Zhizhong Huang** is a senior engineer and lab director in the Institute of Building Performance & Technology, Sino-German College of Applied Sciences, Tongji University, Shanghai, China. **Xiaowei Zheng** is a master's student in the College of Mechanical Engineering, Tongji University, Shanghai, China.

©2009 ASHRAE

the building, such as the heating load as a function of the outdoor temperature. The method was to use the range between the 1st and the 3rd quartile of the quantile–quantile (q–q) plot to check if the heating losses and the outdoor temperatures have the same distribution and, if yes, to perform the regression in this range of the q–q plot. The result was a model that conserves its prediction performance for data sets of the outdoor temperature different than those used for parameter identification. Pedersen (2007) provided an overview of the background for meteorological and sociological influences on thermal load and energy estimations. As his point of view, "regression analysis is mainly based on large amounts of metered load data, long-term weather characteristics and some information about the buildings being modeled. A statistical approach is most suitable for large development areas and long-term estimates of the expected load and energy demand." Freire (2008) adopted independent variables – heating, ventilation and air conditioning (HVAC) power, outdoor temperature, relative humidity and total solar radiation – to obtain the regression equations that were used to define a couple of linear Multiple-Input/Single-Output (MISO) models, since two main outputs were involved, indoor temperature and relative humidity. And validation procedures have shown very good agreement between the regression equations and the simulation tool for both winter and summer periods.

However in China there are few researchers working in this field and few accomplishments have been achieved. In this paper the authors explore the data in the energy consumption database of commercial buildings in Shanghai by data mining techniques to find the relationship among building energy consumption and such relative variables as building area, function of building, cooling/heating sources, etc, so as to establish a regression model for commercial building energy consumption estimation in Shanghai.

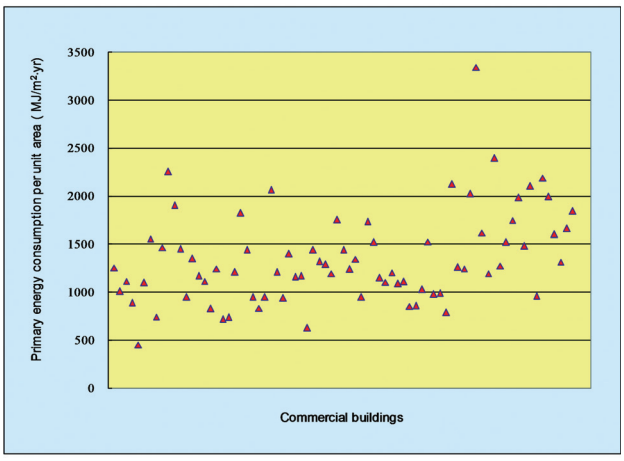

Figure 1 *Annual primary energy consumption of 77 commercial buildings in Shanghai.*

A BRIEF INTRODUCTION TO THE DATABASE

Commercial buildings include office buildings, hotels, shopping malls, and those buildings which contain office, hotel, retail, entertainment, restaurant, etc. The energy consumed by commercial buildings accounts for about 1/3 of the total building energy consumption in China (Tu and Wang 2004). According to a statistical result, energy consumption per unit area of normal commercial buildings is about 5 times of that of normal residential buildings, and that of Grade-A commercial buildings reaches 15 to 20 times of that of normal residential buildings (Lang 2005). Thus, an energy consumption database of commercial buildings in Shanghai is needed for investigation, statistics and analysis of a significant part of the energy consumption of Shanghai.

There are two types of data in the database. One type refers to basic descriptive data of each building such as the owner, location, year built, total building area and the area ratios of office, commercial and hotel parts, characteristics of envelope, types of HVAC equipment and system, etc. The other type is the energy consumption data of the building, i.e., the monthly energy consumption of each building, including electrical power, gas, oil and coal use, and the total primary energy consumption. Up to now, the database has contained 95 commercial buildings in Shanghai, and more buildings' data will be collected in the future. Figure 1 illustrates the annual primary energy consumption per unit area of 77 commercial buildings. There are 18 buildings left out because of the missing of energy consumption data.

DATA MINING PROCESS

The purpose of data mining of the energy consumption database of commercial buildings in Shanghai is to find the relationship between annual primary energy consumption of buildings and other relative variables, and then to establish an energy consumption model for commercial buildings in Shanghai. Since the 77 commercial buildings are randomly sampled, the model developed from them is good for representing commercial buildings in Shanghai. The process of data mining mainly includes the phases of sampling, exploring, modifying, modeling and assessing, as shown in Figure 2.

In the sampling phase the data is sampled by extracting a portion of a large data set which is big enough to contain the significant information, yet small enough to manipulate quickly. Mining a representative sample instead of the whole volume reduces the processing time required to get crucial business information.Then the sampled data is explored by searching for unanticipated trends and anomalies in order to gain understanding and ideas. Exploration helps refine the discovery process. If visual exploration doesn't reveal clear trends, people can explore the data through statistical techniques including factor analysis, correspondence analysis, and clustering. After the exploration phase, data is modified by creating, selecting, and transforming the variables to focus on the model selection process. Based on the discoveries in the exploration phase, people may need to manipulate the data to

include information such as the grouping of customers and significant subgroups, or to introduce new variables. During the modeling phase, software searches the data automatically for a combination of data that reliably predicts a desired outcome. Modeling techniques in data mining include neural networks, tree-based models, logistic models, and other statistical models -- such as time series analysis, memory-based reasoning, and principal components. Once modeling is finished, the model obtained should be assessed by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs. A common means of assessing a model is to apply it to a portion of data set aside during the sampling phase. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model.

Definition of Variables

After a preliminary search, 14 variables are selected for data mining; they are building age, gross building area, area ratio of office to gross area, area ratio of retail to gross area, area ratio of hotel to gross area, U-value of windows, window glass tinted, cooling capacity, cooling primary energy efficiency, heating capacity, heating primary energy efficiency, building automation system, annual operation time of HVAC system, and total primary energy consumption, as listed in Table 1.

In the process of data mining, all data are divided into training sample sets, verification sample sets and test sample sets. Since the database has not contained abundant data up to now, all data in it are adopted as training sample sets in order to achieve a better data mining result.

Handling of Missing Data

Some samples contain missing variables, and the worst is variable FILM. The presence and absence of each variable are shown in Table 2.

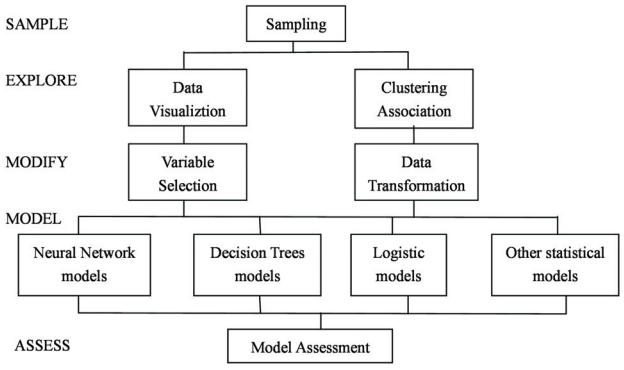

Figure 2 *Flow chart of data mining.*

In statistics the usual methods to deal with missing data include mean/mode attribution method, regression attribution method and multiple attribution method (Grzymala-Busse and Hu 2000). Among the three methods, mean/mode attribution method and multiple attribution method will make remarkable interpolation errors under small sample condition, which will have influence on the data mining results. And there is a functional relationship between the cooling/heating capacities and the other variables. For example, it is obvious that the larger the building area the larger the cooling/heating capacities; will be the better the heat insulation capability of the windows, the smaller the cooling/heating capacities will be. Therefore, the regression attribution method is selected to interpolate the missing data if the data cannot be obtained by other means.

After the authors' efforts via further questionnaire and on-site investigation to find and fetch missing data, only two variables – HCAPACITY and ENERGY, still contain missing data (5 and 15 missing points respectively). The regression attribution method is adopted to establish a regression model for the missing HCAPACITY data. All the samples missing primary energy consumption data (variable ENERGY) are excluded from the regression analysis, because it is the dependent variable of the final building energy consumption regression model.

A regression analysis is conducted on a data set containing the 80 samples having complete data. It is supposed to be a non-linear relationship between the dependent variable HCAPACITY and the other independent variables. Step-wise regression, principal component analysis, and Mallow's Cp statistic method are used respectively for variable sieving. Since the result obtained from step-wise regression has the least F-test value and t-test value, this method is selected finally. The F-test value and t-test value of the model are all less than 0.15, which satisfies the significance level demand of 5%. The adjusted R^2 value is 0.9947. The regression equation found for the missing values of HCAPACITY is:

$$\ln(HCAPACITY) = 0.737\ln(AREA) + 0.059(WINU) + 0.822(RETAIL) \quad (1)$$

The independent variables of the regression equation are AREA, WINU, and RETAIL. In general, the larger the gross building area, the larger the heating load is. The U-value of window has impact on the heat gain of building, and the larger the U-value, the greater the heat loss rate is to the outdoors. The retail part in a building consumes more heating energy than the office and hotel parts, so the larger the area ratio of retail, the more heating energy is needed. Therefore, there exist positive correlations between HCAPACITY and each of the other variables, respectively.

The regression equation, Equation (1), is used to predict the heating capacity of the 89 buildings in the database with actual heating capacity data; the predicted result and actual data are compared and the relative errors between them are calculated as shown in Figure 3. The figure shows that the relative errors are mostly located in the range of ±20%.

Table 1. Variables for Data Mining

Variable	Name	Data type	Data	Units
Building age	BUILTTIME	Numeric, continuous	The deadline is the end of 2004	
Gross building area	AREA	Numeric, continuous		sq.m.
U-value of windows	WINU	Classified	6.4 for single glazing, and 3.2 for double glazing	$W/(m^2 \cdot K)$
Window glass tinted	FILM	Boolean	FALSE means untinted, TRUE means tinted	
Area ratio of office to gross area	OFFICE	Numeric, continuous		
Area ratio of retail to gross area	RETAIL	Numeric, continuous	Fractions of total area. The sum of the three is equal to or less than 1.	
Area ratio of hotel to gross area	HOTEL	Numeric, continuous		
Cooling capacity	CCAPACITY	Numeric, continuous	The sum of cooling capacities of all chillers	kW
Cooling primary energy efficiency	CPER	Numeric, continuous	Calculated based on primary energy. The COP of chillers used are: reciprocating: 3.75; screw: 5.1; centrifugal: 5.13; heat pump: 2.8; direct fired absorption: 1.17**	
Heating capacity	HCAPACITY	Numeric, continuous	The sum of the heating capacities of all types of heating source	kW
Heating primary energy efficiency	HPER	Numeric, continuous	Calculated based on primary energy. The thermal efficiencies used are: oil boiler: 90%; gas boiler: 90%; coal boiler: 78%; direct fired absorption chiller: 90%; heat pump: 3.4 (COP)	
Building automation system	BAS	Boolean	FALSE means not installed, TRUE means installed	
Number of Operation hours of HVAC system per day	ACTIME	Numeric, continuous		Hour
Annual primary energy consumption	ENERGY	Numeric, continuous	Total energy consumption of building, converted into primary energy*	MJ

*1kWh electricity = 11.5 MJ primary energy
**These numbers are approximations for each type of equipment. The actual COP of a specific chiller in a specific building depends on the chiller model and its condition.

Table 2. Missing Variables

Variable	BUILTTIME	AREA	WINU	FILM	OFFICE	RETAIL	HOTEL
Subsistent	89	94	69	40	84	85	89
Missing	6	1	26	55	11	10	6
Variable	**CCAPACITY**	**CPER**	**HCAPACITY**	**HPER**	**BAS**	**ACTIME**	**ENERGY**
Subsistent	88	89	84	86	95	72	79
Missing	7	6	11	9	0	23	16

Considering the limited quantity of data in the database, this error distribution is acceptable.

Test of Extraordinary Points

The methods for extraordinary points tests provided by SAS are Cook distance statistics (COOKD) and studentized residual elimination (SRE(i)). Generally, a measured value is considered an extraordinary point if its COOKD>50% or |SRE(i)|>3. According to this principle, there are only two samples with |SRE(i)| bigger than 3. Among them, Building A has the cooling capacity per area twice that of the other buildings with comparable building area, because it is mainly a shopping mall, while most of the other buildings are mainly office and hotel. The annual primary energy consumption per unit area of Building B reached 7546.3MJ/(m^2·yr), which is extraordinary high without reasonable explanation. The authors think these two samples are extraordinary and should be eliminated from the data sets.

Data Mining

There are various data mining algorithms provided by SAS, such as decision tree method, generic algorithm, and so on. Considering there are only 95 samples in the database, and statistical methods are usually helpful to achieve better results under small sample condition, they are selected for data mining and the establishment of a building energy consumption model.

After interpolating for missing data and eliminating extraordinary data, there are 93 samples left for data mining. Since the final objective is to establish a model of building primary energy consumption per unit area per year, a new variable – EUI, is introduced, defined as,

$$EUI = \frac{ENERGY}{AREA \cdot YEAR} \tag{2}$$

EUI is used as the new dependent variable of a regression model, and the other variables, except ENERGY and AREA, are all defined as independent variables.

Although the statistical methods are selected for data mining, a further selection is necessary to decide which method is the most suitable among forward selection, backward elimination, step-wise regression and principal components analysis. Two judging criteria are used for the method selection: one is that the F-test value and t-test value of the model should be less than a significant level of 5%; the other is that after variable reduction, all variables finally contained in the model and the sign of their coefficients should be consistent with physical expectations based on relevant professional knowledge.

Four models were established by using forward selection, backward elimination, step-wise regression and principal components analysis. The models are compared according to the criteria described above. The model established by step-wise regression best met the criteria and is the final selection. The F-test and t-test values of this model are all less than 0.15, satisfying the significance level of 5%. The adjusted R^2 value is 0.9081 and the variables contained in the model are ACTIME?CCAPACITY?OFFICE and HOTEL. The resulting regression equation is:

$$EUI = 36.704(ACTIME) + 0.040(CCAPACITY) \\ 710.7(OFFICE) + 1108.6(HOTEL) \tag{3}$$

The T-scores of variables (shown in Figure 4) show that CCAPACITY is the most significant variable to EUI, followed by OFFICE, ACTIME and HOTEL in turn.

Model Evaluation

As mentioned above, every sample has 14 variables, but only 4 of them are finally selected as variables of the model. Among these 4 variables, the difference of coefficients of

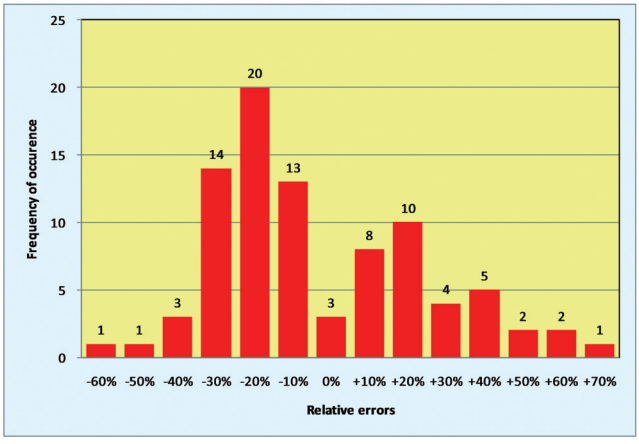

Figure 3 *Relative errors of predicted values of heating capacity.*

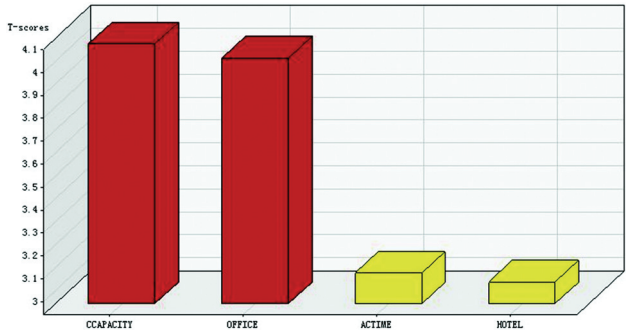

Figure 4 *T-Scores of Variables.*

OFFICE and HOTEL shows the influence of different building functions. Because the HVAC operation times of hotels are normally longer than those of offices, and a great amount of energy is consumed by various facilities and appliances in hotels, this leads to a larger coefficient to the variable HOTEL than OFFICE. For variable RETAIL, certainly a retail building consumes more energy than an office building, but in the database there are few data for retail buildings or buildings with large retail area. So it is reasonable to eliminate this variable because its explanatory significance is low.

There are two reasons why the variable CCAPACITY is selected and HCAPACITY is eliminated. One is that all the buildings in the database are located in Shanghai, where the cooling load is larger than the heating load and the annual cooling degree-days are much more than heating degree-days. The other is that almost all of these buildings contain very large internal zones that need cooling all year round. These two reasons cause cooling to consume much more energy than heating in the buildings in the database, so that the variables HCAPACITY and HPER are eliminated from the model.

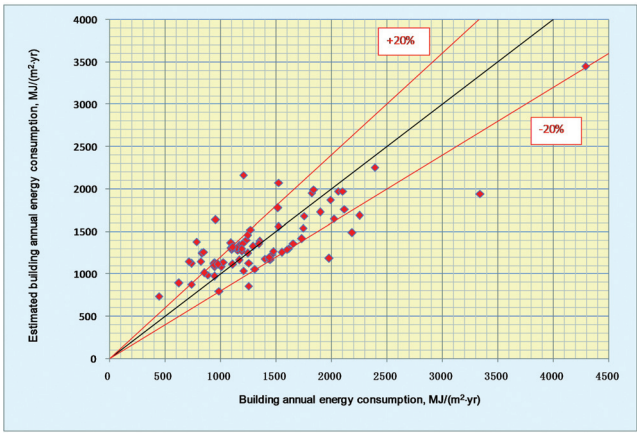

Figure 5 *Comparison of predicted values and actual values of building energy consumption.*

The energy consumed by HVAC system takes account for a big portion (30% to 40% or more) in total building energy consumption, therefore, the variable ACTIME is selected and has a positive correlation with EUI.

The other eliminated variables such as WINU and FILM, together with the heat transfer features of exterior walls, which is eliminated in the preliminary selection of variables, did have impact on the cooling/heating loads in perimeter zones of the buildings, however the energy consumption of these zones accounts for a relatively much smaller portion of the total building energy consumption. Both the variable CCAPACITY and CPER have impact on building energy consumption, however, the regression result excludes CPER. Although there are five types of chillers used in the buildings, the chillers installed in most buildings are centrifugal and screw types, and the COPs of them are almost equal. This leads to the efficiency of the chillers not significantly influencing the model of building energy consumption.

The regression model is used to predict the building energy consumption of each building in the database, and the predicted data are compared with the actual data with the error distribution shown in Figure 5. The most errors are within a range of ±20%, which is an acceptable result.

Two buildings selected randomly outside the training sample sets are also adopted to verify the accuracy of the regression model. The profiles of the two buildings are given in Table 3. The verification results are presented in Table 4. The errors between actual EUI values and those calculated by the regression model are less than 20%, which is also an acceptable result.

CONCLUSION

This paper introduces a data analysis and data mining process applied to the energy consumption database of commercial buildings in Shanghai to find the relationships among annual building energy consumption and various variables. Several data analysis methods and data mining techniques are employed and the results are analyzed. Criteria for comparing regression models – statistical values such

Table 3. Two Buildings Used for Verification of the Regression Model

Building	Total Building Area (sq.m.)	Office Area (sq.m.)	Hotel Area (sq.m.)	Cooling Capacity (kW)	Operation Period	Annual Average Electrical Consumption (kWh)
A	15,381	13,380	0	510	6:30 ~ 17:30	1,846,215 kWh
B	30,000	21,000	0	3,800	8:00 ~ 18:00	3,686,600 kWh

Table 4. Verification Results

Building	AREA (m^2)	ACTIME (hours)	CCAPACITY (kW)	OFFICE	HOTEL	Actual EUI (MJ/ m^2·yr)	Calculated EUI (MJ/ m^2·yr)	Error (%)
A	15,381	11	510	0.87	0	1206.6	1042.5	-13.6
B	30,000	10	3,800	0.7	0	1235.3	1016.5	-17.7

as F-test value, t-test value and R^2 value and correlative professional knowledge are used. During the process of data mining, the non-linear regression method is used for missing data interpolation. After a comparison of regression results according to the criteria mentioned above, the step-wise method is selected as the most suitable method. The regression equation obtained has an R^2 value of 0.9947.

After the missing data interpolation, different data mining methods are used to establish a building energy consumption regression model. Based on the results of the analysis, the step-wise regression method was finally selected, and the regression model of the annual primary energy consumption per unit area for commercial buildings in Shanghai is obtained. Four variables are sieved out from 14, i.e., operation time of HVAC system, cooling capacity, area ratio of office and area ratio of hotel, for the regression model. These variables are the most impactive factors on energy consumption of commercial buildings in Shanghai. The R^2 value of model is 0.9081. The regression model is used to predict the building energy consumption of each building in database, with most errors (78%) within a range of ±20%. Two buildings not contained in the database are also used for verification of the model, which shows acceptable errors of less than 20%. It can be concluded that the regression model with four variables is accurate enough to predict the energy consumption of commercial buildings in Shanghai. However, if the data of more buildings are collected and put into the training sample sets in the future, the model may be different and contain more variables, at the same time the accuracy will be further improved.

REFERENCES

Abdullatif E. Ben-Nakhi, Mohamed A. Mahmoud, 2004, Cooling load prediction for buildings using general regression neural networks, Energy Conversion and Management 45 (2004), pp. 2127-2141.

Cristian Ghiaus, 2006, Experimental estimation of building energy performance by robust regression, Energy and Buildings, 38 (2006), pp. 582-587

Grzymala-Busse J.W. and Hu M., 2000, A Comparison of Several Approaches to Missing Attribute Values in Data Mining. RSCTC, P340–347

Joseph C. Lam, et al. 1997, Regression analysis of high-rise fully air-conditioned office buildings, Energy and Buildings, 26 (1997), pp. 189-197.

Lang Siwei, 2005, The key of the energy saving design code for public buildings, Construction Science and Technology, No. 13, 2005.

Linda Pederson, 2007, Use of different methodologies for thermal load and energy estimations in buildings including meteorological and sociological input parameters, Renewable and Sustainable Energy Reviews, 11 (2007), pp. 998-1007

Roberto Z. Freire, et al. 2008, Development of regression equations for predicting energy and hygrothermal performance of buildings, Energy and Buildings, 40 (2008), pp. 810-820.

Simulation Research Group, Lawrence Berkeley National Lab, 1991, DOE-2 Basics.

Tu fengxiang and Wang Qingyi, 2004, Building energy saving --- A certain selection of China energy saving stratagem (I), Energy and Environmental Protection, No. 8, 2004.

William Chung, et al., 2006, Benchmarking the energy efficiency of commercial buildings, Applied Energy 83 (2006), pp. 1-14.

Copyright of ASHRAE Transactions is the property of American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.