The 2005 World Sustainable Building Conference, Tokyo, 27-29 September 2005 (SB05Tokyo)

AN OVERVIEW OF UTILITY BILLS ANALYSIS METHODS FOR PREDICTING BUILDING ENERGY CONSUMPTION IN TROPICAL REGION

Bing Dong[1]
Siew Eang Lee[2]

[1] Energy and Sustainability Unit, School of Design and Environmental, National University of Singapore, 117566, Singapore, bingdong@nus.edu.sg
[2] Energy and Sustainability Unit, School of Design and Environmental, National University of Singapore, 117566, Singapore, bdgleese@nus.edu.sg

Keywords: utility bills, regression analysis, weather data, Neural Networks, Support Vector Machines, Singapore

Summary

This paper reviews several utility bills analysis methods for predicting building energy consumption in the tropical region. Firstly, a multiple linear regression analysis method is applied to six commercial buildings, trying to establish an accurate linear prediction models. Three independent variables, namely, outdoor dry-bulb temperature (T_0), relative humidity (RH) and global solar radiation (GSR) are taken as independent variables. Secondly, in order to explore the non-linear performance of landlord energy use, artificial neural networks (NN) and support vector machines (SVM) are utilized to predict landlord energy consumption. Another four buildings are involved in these two methods. The data from four years' bills are used to train and test the models. In addition, R^2, mean absolute error (MAE) and coefficient of variance (CV) are taken as three model performance criteria.

The results show that in most cases outdoor dry-bulb temperature accounts for more than 80% of the changes of whole building energy consumption. In addition, most of the CVs of prediction models are below 7%. Support vector machines present the best prediction results with lowest mean absolute error (MAE) within 4% and CV within 3%. This study is important for energy services companies and building owners to track energy use during the building retrofitting period in the tropical region.

1. Introduction

Since the energy crisis of 1970s, people began to recognise the needs for efficient energy use. In the shortage of natural resources fundamental to the generation of electricity in Singapore, energy remains a critical factor for the success of all economies in the immediate and long-term future. Previous building energy research conducted by the Building and Construction Authority of Singapore (BCA) shows that the energy consumptions in buildings accounts for approximately 37% of the whole electricity consumption in Singapore. Particularly, energy consumption in office buildings accounts for nearly 57% of the total electricity consumption in buildings. Such high energy consumption may be resulted from the combined impact of common deficiencies in building design and operation, for example, outdated and inefficient equipment, improper equipment selection and installation, lack of inadequate commissioning efforts, and inadequate maintenance. One of the cheapest and most useful ways to reduce this high consumption is by enhancing energy efficiency through the application of applying energy conservation measures (ECMs).

A crucial element in the implementation of an energy conservation program is the ability to verify savings from measured energy use data (Fels and Keating, 1993). The determination of energy savings requires both accurate measurement and reliable methodology. However, there is no direct way of measuring energy use or demand savings since instruments cannot measure the absence of energy use or demand after retrofitting. The baseline model developed from utility bills provides a way to compute such savings. The accurate of the baseline model is totally based on the accurate of prediction methods. Many efforts have been taken on the development and improvement of the prediction methods as accurate as possible (Fels *et al.,* 1986; Kissock *et al.*, 1993; Krarti *et al.,* 1998; Dhar *et al.,* 1999). However, in most practical cases, utility bill data are used because they are widely available and inexpensive to obtain and process. In the tropical region, B.Dong *et al.* (2005) utilized two years' utility bills to establish a baseline model and the results showed MAE.

In this paper, a review of utility bill analysis methods in the tropical region was presented based both on the whole building level and landlord level. The reason for the research on the landlord energy consumption is that the building owner often received the landlord bills only. It seems more meaningful to baseline landlord energy consumption rather than the whole building energy use for the benefits of both building owners and ESCOs. However, the method on the whole building energy consumption will be evaluated firstly. Totally ten

buildings are involved in this study. In all the methods applied, the weather parameters, namely, outdoor dry-bulb temperature (T_0), relative humidity (RH) and global solar radiation (GSR) are taken as three independent variables.

2. Performance Criteria

The criterion used to select the most appropriate regression model is to maximize the goodness of fit using the simplest mode or combination of models (Draper and Smith, 1981). According to the literature review, it is believed that the coefficient of determination (R^2) and the coefficient of variance of the root-mean-square error (CV-RMSE) are two major measures to evaluate the goodness of fit of a model. The CV-RMSE is a non-dimensional measure that is found by diving root-mean-square error (RMSE) by the mean value of total energy consumption E . It is usually presented as a percentage. A CV-RMSE value of 10% indicates that the mean variation in E not explained by the regression model is only 10% of the mean value of E (Reddy *et al.,* 1997a). The CV-RMSE defined below:

$$CV - RMSE = \frac{RMSE}{\overline{E}}.100 \tag{1}$$

Where,

$$RMSE = \left[MSE\right]^{1/2} = \left[\frac{\sum_{i=1}^{n}\left(E_i - \hat{E}\right)^2}{n-p}\right]^{1/2} \tag{2}$$

E_i is the measured energy consumption of single month/day/hour i (i=1,…n). $\overline{E}$ is the mean value of E . $\hat{E}$ is the value of E predicted by the regression model, n is the number of observations; p is the number of model parameters.

For simplicity, the direct deviation between the two types of energy consumption values, measured versus predicted, which is also called mean absolute error (MAE), is defined below:

$$MAE(\%) = \frac{\overline{E}_{predicted} - \overline{E}_{measured}}{\overline{E}_{measured}} \times 100 \tag{3}$$

How to determine the goodness of the model with the two measures, namely, R^2 and CV-RMSE, has been well discussed and examined by researchers. Fels *et al.* (1993) arbitrarily suggested that monthly models with $R^2 \geq 0.7$ and CV-RMSE ≤ 7% be deemed “good” models. Reddy *et al.* (1997a) pointed out that CV-RMSE less than 5% are considered excellent models, those less than 10% are considered good models, and those less than 20% are taken to be mediocre models and those greater than 20% are considered to be poor model. In this study, the values of R^2 and CV-RMSE are following the criteria pointed out by Reddy *et al.* (1997a) because it was concluded from the whole building energy consumptions.

Another important evaluating parameter for model prediction ability is the variance of forecast error (VAR). VAR is defined below (Tan, 2002):

$$VAR(\hat{E} - E) = S^2\left[1 + \frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right] \tag{4}$$

Where, S is the standard error of the estimate as given below:

$$S = \sqrt{\frac{\sum e_i^{\,2}}{n-p}} \tag{5}$$

And, e_i is value of the residual; n is the sample number; p is the model parameter. Finally, the predicted energy consumption may be expressed as $E = \hat{E} \pm \sqrt{VAR}$.

3. Approaches

3.1 Possibilities of Utility Bill Reading Dates

Before any approach is concerned, an important thing in the model development is to verify uncertainty in the bill reading dates. Neither the building owners nor Singapore Power Service Company provide detailed information on utility bill reading dates. In addition, every building may have its own policy for recording electricity use. Hence, it is difficult to verify the specific utility reading period. Here, the methodology was recommended by Reddy et al. (1997a) who pointed out that there were three possibilities in recording building energy use:

a) The utility bill period is correspondent with weather data period. It means that the utility bill reading dates are the same dates as weather data dates.

b) The utility bill period is one month later than the weather data period. It means that the present utility bill actually represents the previous month's electricity use. This situation always happens

c) The utility bill period is 15 days later than the weather data period. It means the readings of utility bills begin around the middle of the month.

For each of the baseline model, all these three possibilities are performed. Then, the one which has the best-fit regression is selected.

3.2 Whole Building Energy Consumption

3.2.1 Data collection

Six buildings were selected randomly among all the buildings around the Central Business District in Singapore. They are all office buildings for commercial use. The utility bills of these six buildings were collected through surveys which were carried by the previous research on building efficiency (Lee, 2001). In order to retain the individual building anonymity, these twelve buildings are referred to as building A, B, C, D, E and F. Table 1 shows the building size and the annual energy use of these buildings. For building A, B and C, 2000 was their modeling year. Building D, E, H, I and J take year 2001 as their modeling year. The utility bills from building F are only available from September 1999 to September 2001 and therefore the baseline year is set to be from October 2000 to September 2001.

Table 1 Size and energy use in six buildings

Building	Modeling Year	Total Bldg. Area(m^2)	Air Conditioned Area(m^2)	Total Bldg. Energy Consumption (MWh/yr)
A	2000	20 165	12 268	3 470
B	2000	32 368	24 825	6 034
C	2000	42 026	25 822	9 998
D	2001	60 894	36 688	11 551
E	2001	42 060	25 833	8 031
F	Oct 00~ Sept 01	43 187	34 753	7 880

2.2.2 Multivariate regression analysis (MLR)

The multiple linear regression model is derived as follows:

$$\hat{E} = \beta_0 + \beta_1 T_0 + \beta_2 RH + \beta_3 GSR \quad (6)$$

Where, β_1, β_2 and β_3 are regression coefficients.

An important verification parameter in the multiple linear regression (MLR) analysis is the partial correlations. The partial correlations procedure computes partial correlation coefficients that describe the linear relationship between two variables while controlling for the effects of one or more additional variables (SPSS 1999). This parameter clearly shows the correlations between every independent variable with the dependent variable (Dong et al. 2005).

The whole multiple linear regression analysis is processed using backward elimination method in SPSS. After the regression model containing all variables has been set up, the partial F-test is calculated for every predictor variable. Based on the critical F-value, which are defined as 0.1 in this study, the backward procedure removes all unneeded X-variables one by one. This backward method lists all possible R^2 based on three possible variables. Hence, when adding additional variables, it can check its effect to the regression model itself. The detailed process can be referred to Dong et al. (2005).

3.2 Landlord Energy Consumption

A buildings' landlord energy consumption refers to the energy utilized by the common facilities, systems, services and space provided by the landlord. typically comprising: a) Air-conditioner central plant system; b) Vertical transportation service such as escalator and lift; c) Ventilation system such as exhaust fan and ventilator; d) Artificial lighting system in the common area. Obviously, the usages of these systems present certain non-linear performance between building energy use and weather data.

3.2.1 Data Collection

Another four buildings, namely G, H, I and J were selected and, this time, four years' bills were collected. The details are shown in Table 2.

Table 2 Size and energy use in four buildings

Building	Training Year	Test Year	Building Design Efficiency (%)	Total Landlord Energy Consumption (MWh/yr)
G	Oct.1996~Oct. 1998 and 2000	2001	61.44	5,291
H	Oct.1996~Oct. 1998 and 2000	2001	56.26	6,024
I	Oct.1996~Oct. 1998 and 2000	2001	54.77	7,681
J	Oct.1996~Oct. 1998 and 2000	2001	77.69	15,400

3.2.2 Neural Networks (NN)

The neural networks toolbox in this study is MATLAB 6.5. The algorithm applied inside is back-propagation (BP). The network had one input layer, one hidden layer of different neurons and one output layer. The transfer functions for the three layers were tan-sigmoid, tan-sigmoid and linear, respectively.

The total input parameters are four including three weather data and the time tag. The output is the building landlord's energy consumption. The first three years' data was used for training and another one year utility bill was used for testing and verification. In this study, the projected year was selected forward to year 2001. It is the same when someone wants to predict backward. The same input information was collected in year 2001. Finally, the predicted annual whole building landlord energy consumption was compared with the actual measured value.

In the training of the back-propagation network, the number of neurons in the hidden layer is set from 1 to 10 because the size of the network should be controlled by the ratio of free parameters to the number of training samples. By varying the number of hidden nodes, the best performance of BP networks was determined based on the MSE. The number of epochs is set to be 500 in this study. The stop criterion is set to be 10^{-5}. After the number of hidden nodes of best performance is chosen, neural networks ran 30 times again on the optimum point and the averages of best five results were selected. It is because the random selection characters of neural networks.

3.2.2 Support Vector Machines (SVMs)

Support vector machines (SVMs), developed by Vapnik and his co-workers in 1995, has been widely applied in classification, forecasting and regression of random data set. One of its main application fields in regression modeling is the time series financial forecasting. Their practical success can be attributed to solid theoretical foundations based on Vapnik-Chervonenkis Theory (Cherkassky *et al.*, 2004). The detailed theory can be found in Vapnik *et al.* (1996). Simply speaking, SVMs is based on the structural risk minimization (SRM) inductive principle which seeks to minimize an upper bound of the generalization error consisting of the sum of the training error and a confidence level. This is the difference from commonly used empirical risk minimization (ERM) principle which only minimizes the training error. Based on such induction principle, SVMs usually achieves higher generalization performance than the traditional neural networks that implement the ERM principle in solving many data mining problems (Dong et al. 2005).

This is the first time that SVMs is applied in the building energy consumption prediction. The principle of SVMs is to map non-linear functions in the low space to the high space to be linear problems by the use of the kernel function. All necessary computations can be performed directly in a high dimensional feature space and a linear function is trained in such a space, without having to compute the map $\phi(x)$ .Some popular kernel functions are the linear kernel $K(x_i, x_j) = x_i \cdot x_j$, polynomial kernel $K(x_i, x_j) = (x_i \cdot x_j + 1)^d$ and

the radial basis function (RBF) kernel $K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$, γ >0, where d and γ are the kernel parameters. Most of the previous research selected Gaussian function which is included in RBF as the kernel model for regression. The selection of parameters of the kernel function, namely C and ε is based on the stepwise search pointed out by Dong *et al.* (2005). In the stepwise method, one-time search was first conducted to get MSE_1. Then, the same selection process is conducted again on parameter C (fixing the first result of ε) and ε (fixing the second result of C), to get lowest MSE_2. The one-time search continues until $MSE_n - MSE_{n-1}$<0.00001, which is the normal stop criteria for neural network training, and then, the training stopped. Finally, after the best (C, ε) is found, the whole training set is trained again to generate the final regressor.

Libsvm-2.6(Chang and Lin, 2001) developed by Chih-Chung Chang and Chih-Jen Lin, is applied in this study to produce and test the application of SVMs on predicting building energy consumption. The training inputs and outputs are the same as NN.

4. Results and Discussion

4.1 Model Identification

4.1.1 Whole building approach

Table 3 shows the results of multiple linear regression analysis. This MLR regression process checked the individual effect of each variable to the whole MLR model and settled down the final variable in the projected regression baseline model. As shown in table 3, for example, the partial R of 0.92 for T_0 means that the outdoor dry-bulb temperature explains 92% of the variation in the whole building energy consumption. For all these six buildings, T_0 explains most of the variation in the whole building energy use followed by RH and GSR. In addition, judging from the t test, at 90% confidence level, only T_0 is statistically significant for most of the buildings. RH and GSR are both rejected for all buildings except building A. Hence, at 90% confidence level, only the model for building A is accepted as a multiple linear regression model. All other models for the left five buildings are single variant linear regression models. However, as it is shown that T_0 is the main contribution to the changes of whole building energy consumption, namely, all the partial R of T_0 is more than 80%, and to be simpler and more practical, in this paper, we only consider T_0 as our model independent variable.

Table 3. Summary of Multiple Linear Regression Results

Variables	A			B			C		
	Correlation		t	Correlation		t	Correlation		t
	Partial R	Model R^2		Partial R	Model R^2		Partial R	Model R^2	
T_0	0.93		7.10	0.86		4.77	0.82		4.6
RH	0.35	0.92	1.01	0.36	0.79	1.07	0.15	0.76	0.43
GSR	-0.66		-2.49	-0.35		-1.06	-0.15		-0.43

(Continued)

Variables	D			E			F		
	Correlation		t	Correlation		t	Correlation		t
	Partial R	Model R^2		Partial R^2	Model R^2		Partial R	Model R^2	
T_0	0.87		5.06	0.86		4.8	0.89		5.6
RH	0.12	0.78	0.31	-0.28	0.81	-0.84	0.44	0.82	1.37
GSR	-0.13		-0.37	0.08		0.23	0.18		0.51

After the model was identified, it was used to predict energy consumption for another 12 months. The results of the prediction are shown in table 4. The direct deviations from modeled and measured energy use are all below 5%, except for building D. The lower the deviation, the more accurate the predicted energy consumption is. For building A, the monthly electricity use in year 2001 is predicted to be 24.11(±1.71) (kWh/m^2/month), an increase of about 0.92%, while the measured energy use is 23.52(kWh/m^2/month), an increase of about 2.45%. They are very near. The results in building D are not good enough.

Table 4 Prediction results of whole building energy consumption

Building	Annual Energy Use (kWh/m^2/month)		CV (%)	VAR(Model) (kWh/m^2/month)	MAE (%)
	Measured	Modeled			
A	23.52	24.11	3.48	2.92	2.51%
B	20.26	20.59	3.80	1.35	1.64
C	32.87	32.09	2.78	2.60	2.42
D	28.59	24.22	3.49	29.79	15.33
E	25.55	25.76	3.76	2.05	0.8
F	18.81	18.34	3.32	2.83	2.5

4.1.1 Landlord approach

The same utility bill analysis method was applied when selecting the suitable bill period.

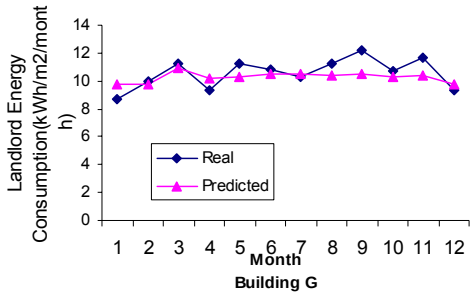
Table 5 Results of neural networks training and prediction

Building Ref. No.	Actual Value ($kWh/month/m^2$)	Neuron Numbers in The Hidden Layer/Predicted Consumption					
		8	9	10	MSE	Prediction CV (%)	MAE (%)
G	10.55	10.39			0.78	9.67	-1.5
H	11.05			11.12	3.83	15.5	0.64
I	9.54			10.23	1.58	15.12	7.33
J	12.59		11.82		2.38	14.17	-6.2

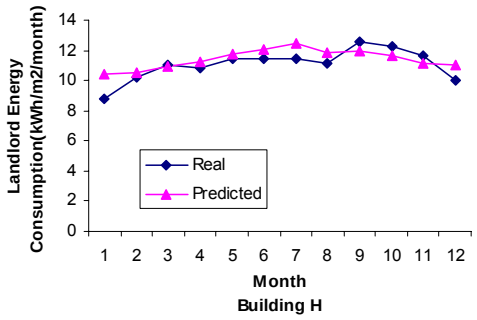
Table 5 shows the results of prediction CV and prediction accuracy. All the prediction CV is larger than 7%, which is deemed the CV of an excellent model. However, the MAE, which show the annual energy consumption prediction error, are all below 10%. Two out of four MAE are below 2%. For building H, it is particularly good, which is only 0.64%. However for building I, both the prediction CV and MAE are highest among these four buildings, which are 15.12% and 7.33 respectively. It shows that the NN model has a good performance in forecasting the annual energy consumption for the landlord energy consumption, but not the monthly consumption. This undesirable result initiated our motivation to search an improved new method to predict the building landlord energy consumption.

The summary of results of SVMs is shown in Table 6. Table 6 shows that building G has the highest MSE of 0.73, while building J has the lowest MSE of 0.14. All CVs, which represent the variances from the true value, are very small and those values are less than 3%. This indicates all SVMs models can be considered as excellent models according to Reddy *et al.* (1997a). Comparing with other studies conducted using other methods such as neural networks (NN) and genetic programming (GP) on the building load research based on hourly or daily data, which are 1993 ASHRAE Competition demonstrated the best CV of 10.36% (NN) on the whole building energy consumption, Kreider *et al.* (1998) found the best CV of 4.7% (NN) on chilled water and Chen *et al.* (2003) found the best CV of 14.7% (GP) on the HVAC load, SVM in this study shows better results in terms of CV. The highest CV of 2.89% appears in Building G, while the lowest CV of 0.99% appears in Building J. Furthermore, the MAE is also small. The best MAE appears in building I, which is only 0.68%.

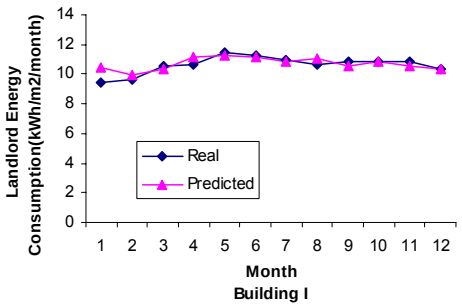
Figure 1 shows the graphical results of predictions for four commercial buildings. Obviously, Building J shows the best prediction result. Because of low CV in building J, the predicted values are almost the same as real values. In addition, all four predicted value curves tracked the variation of real values correctly. It indicates that such kind of method can be applied in tracking the monthly building energy use for diagnosing whether the systems are working properly or not.



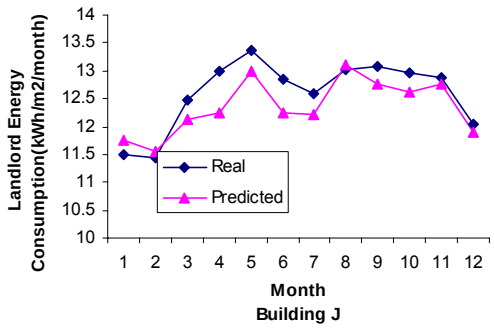
(a)



(b)



(c)



(d)

4.2 Comparison of Three Approaches

Although the buildings for predicting whole building energy consumption are different from the buildings for landlord energy consumption, the prediction results still potentially indicate the differences between these three methods. From table 4 and 5, in terms of MAE, the MLR and SVMs method is slightly better than NNs method. While with regard to the value of CV, SVMs is much better than the other two methods with all value below 3%. Table 5 shows the results of comparison on landlord energy consumption between NN and SVMs. The better CV indicates the stronger tracking ability of SVMs on the monthly basis. It is because of the Structural Risk Minimization (SRM) principle and unique and optimal solution of SVMs. The final results demonstrated that SVMs is feasible and applicable in prediction of monthly landlord utility bills in the tropical region. Moreover, the application of this methodology is not limited to only the tropical region based on its strongly theoretical background and regression characters.

Table 6 Results of comparison on landlord energy consumption between NN and SVMs

	MAE		CV(%)	
Building	NN	SVMs	NN	SVMs
A	-1.5	-2.72	9.67	2.69
B	0.64	3.44	15.5	2.39
C	7.33	0.68	15.12	1.28
D	6.2	-1.89	14.17	0.99

5. Results and Discussion

The methodology for prediction building energy consumption is important for developing the baseline models, which is crucial and necessary for any performance contracting and M&V protocol. The baseline model is a key to secure and verify savings from energy retrofitting programs. The main purpose of this paper is to investigate several mathematical methodologies for prediction and comparison. The weakness and strangeness of different methods are presented as the different prediction results. Such holistic analysis is useful for energy services companies (ESCOes) to select appropriate methods according to different situations such as available bills and model levels. The main contribution of this study is to explore a new method called support vector machines and stepwise search for its parameters' optimization. Finally, the results of baseline models clearly and accurately show and simulate the tendency of energy consumption along the time. They help building owners to track building normal operations and the ESCOes to secure their energy savings in the EPC or IPMVP programs.

References

A. Dhar, T.A. Reddy, D.E. Claridge, A. Fourier, Series model to predict hourly heating and cooling energy use in commercial buildings with outdoor temperature as the only weather variable, J. Solar Energy Eng. 121,1999, 47–53.

B. Dong, S.E. Lee, M.H.Sapar, "A holistic utility bill analysis method for baselining whole commercial building energy consumption in Singapore", Energy and Buildings, Vol. 37 (2), 167-172, 2005

B. Dong, C.Chen, S.E. Lee, "Applying support vector machines to predict building energy consumption in the tropics", Energy and Buildings, vol. 37 (5), 2005.

Chang, C.-C. and C.-J. Lin, 2001, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, Z.Q., Nelson, R.M. and Ashlock, D.A. (2003). Comparsion of methods for predicting monthly psot-retrofit energy use in buildings. ASHRAE Transactions, Vol. 109. pp. 449-459

Draper, N., and H.Smith. 1981. Applied regression analysis, 2nd ed. New Yok: hohn Wiley and Sons.

Fels, M.F., and K.M. Keating. 1993. Measurement of Energy Savings from demand-side management programs in U.S. electric utilities. Annual Review Energy Environ. 18:57-88.

Haberl.S.J. Baseline Calculations for Measurements and Verification of Energy and Demand Savings in a Revolving Loan Program in Texas, ASHRAE Transactions 1998, V. 104, Pt. 2

Kissock, J.K., Reddy, T. A., Claridge D.E., 1998, Ambient-Temperature Regression Analysis For estimating Retrofit Savings in Commercial buildings, Transactions of the ASME, Vol. 120, p168-176.

Krarti, M., Kreider, J., Cohen, D., and Curtiss, P., 1998, Prediction of Energy Saving for Building Retrofits Using Neural Networks, ASME journal of Solar Energy Engineering, Vol. 120, No. 3, Aug.

Lee S.E., 2001, Energy Efficiency of Office Buildings In Singapore, BCA Seminar on Energy Efficiency in Building Design, April 18.

Reddy.T.A., Saman, N.F., Claridge, D.E., Haberl, J.S., Turner, W. D., and Chalifoux, A., 1997, Baselining Methodology for Facility-Level Monthly Energy Use – Part 1: Theoretical Aspects, ASHRAE Transactions 1997, v. 103, Pt. 2

Tan, W., 2002, Practical Research Methods, Singapore: Prentice Hall.

V.Cherkassky, Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, Neural Networks, Vol. 17, pp. 113-126, 2004.

V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.