Jurnal Teknologi

Full Paper

DATA SELECTION TEST METHOD FOR BETTER PREDICTION OF BUILDING ELECTRICITY CONSUMPTION

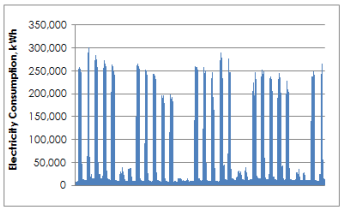
Iqbal Faridian Syah[a,b,c], Md Pauzi Abdullah[a,b*], Husna Syadli[a,b,c], Mohammad Yusri Hassan[a,b], Faridah Hussin[a,b]

[a]Centre of Electrical Energy Systems (CEES), Institute of Future Energy, Universiti Teknologi Malaysia , 81310 UTM Johor Bahru, Johor, Malaysia
[b]Faculty of Electrical Engineering, Universiti Teknologi Malaysia , 81310 UTM Johor Bahru, Johor, Malaysia
[c]Electrical Department, Faculty of Engineering, Universitas Malikussaleh, 24351 Aceh, Indonesia

Article history
Received
23 June 2015
Received in revised form
17 November 2015
Accepted
10 Jan 2016

*Corresponding author
mpauzi@ utm.my

Graphical abstract



Abstract

The issue of obtaining an accurate prediction of electricity consumption has been widely discussed by many previous works. Various techniques have been used such as statistical method, time-series, heuristic methods and many more. Whatever the technique used, the accuracy of prediction depends on the availability of historical data as well as the proper selection of the data. Even the data is exhaustive; it must be selected so that the prediction accuracy can be improved. This paper presented a test method named Data Selection Test (DST) method that can be used to test the historical data to select the correct data set for prediction. The DST method is demonstrated and tested on practical electricity consumption data of a selected commercial building. Three different prediction methods are used (ie. Moving Average, MA, Exponential Smoothing, ES and Linear Regression, LR) to evaluate the prediction accuracy by using the data set recommended by the DST method.

Keywords: Electricity prediction, data selection, prediction accuracy

1.0 INTRODUCTION

The use of electricity in commercial buildings has been growing rapidly since last decades. An increase in the number of buildings in Malaysia give a great impression to the country's development, but it also increases the use of energy. The statistics record by Energy Commission of Malaysia ('SuruhanjayaTenaga') [1] indicates that 94% of electric power in this country is generated by fossil material. Buildings, which consist of commercial and residential building is using 54% of the total energy usage in the country. Commercial building is the main contributor with 33% while residential building is 21%. HVAC is the biggest electricity consumer in commercial building, followed by lighting, and office equipment as illustrated in Figure 1. Knowledge of the future energy usage will bring great benefits to the maintenance personnel of commercial buildings. For example, the forecasted energy usage patterns will help them in analyzing the building's future energy usage and hence plan the main target for energy preservation.

Due to the diversity and complexity of commercial buildings as well as the random usage pattern, predicting electricity consumption of a building is complicated. Too many research works have been concentrated in proposing new techniques or approaches to improve the accuracy of prediction. It is argued that, whatever the techniques or methods that being used, the accuracy of prediction depends on the availability of historical data as well as the

proper selection of the data. In other word, good data is needed for methods/techniques to predict accurately. This paper presented a test method that can be used to identify the correct data to be used in predicting the future electricity consumption.

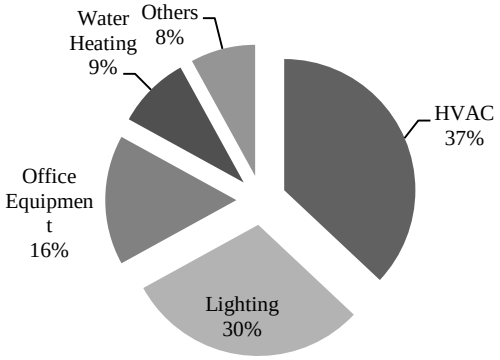


Figure 1 Energy use in commercial buildings

2.0 ELECTRICITY CONSUMPTION PREDICTION

Various prediction methods have been developed in the literature to estimate future electricity consumption either for short term or long term. Due to complexity of the problem, many researchers have applied heuristic methods such as Neural Network, Genetic Algorithm, Support Vector Machines and others.

Weijie Mai et. al [2] proposed Radial Basis Function Neural Network (RBFNN) for predicting hourly energy consumption of large commercial office buildings using outside weather and load data as input historical data. A commercial office building in Shenzhen has been selected for the case study and high accuracy has been demonstrated in the validation method using actual building data under various weather conditions. Kang Ji Li et al [3] used Hybrid Genetic Algorithm-Based Network Adjustment Fuzzy Inference System (ANFIS-GA) and Neural Networks (NN) as comparative studies to predict energy consumption of buildings. Data collected from the energy prediction of library building located at Zhejiang University, China showed that the method gives better performance in terms of accuracy compare to ANN prediction. Languang Zhao et. al [4] predict electricity consumption by using General Regression Neural Network (GRNN) for energy conservation strategy. Electricity consumption data taken from university campus buildings from January 2009 to November 2011 is used. The simulation results show that the prediction accuracy and efficiency are met. In [5], Pedro A. Gonzalez et al present the forecast hourly energy consumption in buildings by using Feedback Artificial Neural Network trained through hybrid algorithms. It is claimed that good and accurate results are achieved. Right BE and U. Teoman [6] applied Back propagation Neural Network (PNN). Heating energy data collected from three different buildings used for prediction. Ahmad Sukri et al [7] proposes Group Method of Data Handling (GMDH) and Least Square Support Vector Machine (SVM LS) to achieve better accuracy in predicting building energy consumption. Bing Dong et al [8] uses Support Vector Machine (SVM) and Neural Networks (NNS) to forecast energy consumption of buildings in tropical regions. Four buildings were used throughout the Central Business District in Singapore. Four building data collected from October 1996 to October 1998 to predict energy consumption. It is reported that the performance of Support Vector Machines (SVM), in terms of CV and MSE is better than using Neural Networks (NNS) and a Genetic Algorithm.

Vladimir Cherkassky et al [9] applied computational intelligence techniques to forecast electricity consumption. The proposed approach combines the regression and classification methods, as a function of time (days) and temperature, using historical data from a number of commercial and government buildings to improve the accuracy of prediction. Empirical comparison shows that the proposed approach provides better improvements.

3.0 STATISTICAL METHOD IN PREDICTION

3.1 Statistics Method for Predicting Electricity Consumption

Statistics play an important role in research. It provides simple techniques in classifying the data and presenting the data more easily, so that data can be more easily understood. Statistics can help researchers to conclude whether a difference obtained was significant. Whether the conclusions drawn representative enough to give inference against certain populations. Statistical techniques can also be used in testing hypothesis, given the purpose of research in general is to test the hypotheses that have been formulated, the statistic can assists researchers in the decision to accept or reject a hypothesis. Therefore, in predicting electricity consumption, statistical methods such as; Moving Average (MA), Exponential Smoothing (ES) and Linear Regression (LR) are widely accepted.

3.1.1 Moving Average (MA) Method [10]

Moving Average (MA) method is built by calculating the average running error generated at any point in time. Generally, the value of the weighted average. Moving Average (MA) method has the form:

$$Y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ \ldots + \theta_q\varepsilon_{t-q} \qquad (1)$$

Where Y_t was estimated value at time t, which is a weighted average of errors in the example earlier time. θvalues is the coefficient term Moving Average. Equation (1) is rewritten as:

$$Y_t = c + \varepsilon_t + \sum_{j=1}^{q} \theta_j\varepsilon_{t-j} \qquad (2)$$

Where q, representing the number of terms of previous errors, known as sequence Moving Average (MA).

3.1.2 Exponential Smoothing Method (ES)[11]

Exponential smoothing (ES) method is a full restore procedure continuing on forecasting new observations on the object. This method gives emphasis on the primacy of the rapid decline in the previous observations of the object.

3.1.3 Linear Regression Method (LR)[12]

Linear Regression (LR) method is statistical method that are used to form the shape of the relationship between the dependent variable (dependent; response; Y) with one or more independent variables (independent, predictors, X). If the number of independent variables there is only one, called a simple linear regression, whereas if there is more than one independent variable, referred to as a multiple linear regression.

Regression analysis has at least three purposes, namely for the purposes of the description of the phenomenon of data or cases that are being tested, for the purpose of control, as well as for prediction purposes. Regression able to describe the phenomenon of data through the establishment of a form of numerical relationships. Regression can also be used to conduct surveillance (control) against a case or things that are being observed through the use of regression models were obtained. In addition, regression method can also be used to perform prediction for the dependent variable. But keep in mind, in the concept of regression prediction can only be done in the data range of independent variables that are used to form the regression method.

3.2 Accuracy of Predictions

Accuracy is one of the fundamental things in prediction, namely how to measure the suitability of a given data set. Accuracy is seen as a rejection criterion for choosing a prediction method. An accurate method givesminimumprediction error whichare commonly measured through the following statistical error measurements; i) Mean Absolute Deviation (MAD), ii) Root Mean Square Error (RMSE), iii) Mean Absolute Percentage Error (MAPE) and iv) Mean Relative Error (MRE).

3.2.1 Mean Absolute Deviation (MAD)

MAD is a measure of the overall forecasting error for a method, which is defined as the mean distinction between the predicted values with the observed data (actual value). This value is calculated by taking the amount of the absolute value (absolute) of any forecasting errors divided by the amount of data period. It can be calculated as follows:

$$MAD = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \quad (3)$$

Where: y_t = actual value

$\hat{y}_t$ = predicted/forecasted value

n = total period

3.2.2 Root Mean Square Error (RMSE)

RMSE is the average squared distinction between the predicted values with observed data (actual value). Weakness use of RMSE is that it tends to accentuate the distinction of great value because of the squaring. For example, when the forecasting error for the period 1 two times greater than the error period 2, then the square error at first period four times greater than the squared error in period 2. Therefore, using the RMSE as forecasting error calculations typically indicate where better distinction has some value smaller than the distinction of great value. It is calculated as follows;

$$RMSE = \sqrt{\sum_{t=1}^{n}\frac{(y_t-\hat{y}_t)^2}{n}} \quad (4)$$

3.2.3 Mean Absolute Percentage Error (MAPE)

MAPE is Average absolute distinction (absolute) between the predicted value with the actual value, expressed as a percentage of the actual value.

$$MAPE = \{\sum_{t=1}^{n}|y_t - \hat{y}_t|\}x100/n \quad (5)$$

3.2.4 Mean Relative Error (MRE)

MRE measures the difference between actual and estimated value relative to the actual value. It is defined as the ration of mean absolute error to the mean value of the measured quantity.

$$MRE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t-\hat{y}_t}{\hat{y}_t}\right|x100\% \quad (6)$$

4.0 DATA SELECTION TEST METHOD

To improve the accuracy of prediction, the historical data should be analyzed first. This paper presents a method for selecting data to assist users in choosing the correct data.

Assume that electricity consumption data of a building is represented by matrix [X] with size $m \times n$ as follows;

$$[X] = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} \quad (7)$$

Where;

i: i^{th} hour, j=1,2…. m^{th} hour

j: j^{th} day, i=1,2…. n^{th} day

$x_{i,j}$: is an element of data matrix, x representing the electricity consumption at i^{th} hour and j^{th} day

The total electricity consumption in one day is represented by matrix [y] with size $1 \times n$ as follows

$$[Y] = [y_1 \quad \cdots \quad y_n] \tag{8}$$

Where each element of matrix [Y] represents the total electricity consumption in one day and is calculated as:

$$y_n = \sum_{i=1}^{m} x_{i,n} \tag{9}$$

The mean and standard deviation of the electricity consumption for m days is given by equation (6) and (7) espectively.

$$\bar{y} = \frac{1}{m}\sum_{n=1}^{m} y_n \tag{10}$$

$$\sigma = \sqrt{\frac{1}{m}\sum_{n=1}^{m}(y_n - \overline{y_n})^2} \tag{11}$$

All historical data is important for predicting future electricity consumption of a building. However, some data has significantly different pattern from the rest and can affect the accuracy of the results. In this paper, this data is called '*Not Useful (NU) data*' and must be removed from the data set. For example, the electricity consumption pattern of a building on Saturday and Sunday (weekend) is significantly different from Monday to Friday (weekdays). Therefore, data on Saturday and Sunday is considered as *NU* data for predicting future consumption on weekdays.

To identify *NU* data from a [X] matrix, the following test is proposed.

Step 1: Identify *NU* data

If p[th] day is *Not Useful(NU)* then the. p^{th} column elements of matrix [X] matrix is called;

$$[X_{day}^{NU}] = \begin{bmatrix} x_{1,p} \\ \vdots \\ x_{m,p} \end{bmatrix}$$

If more than one day is *Not Useful(NU)*, then matrix $[X_{day}^{NU}]$ is a $(m \times p^{all})$ matrix where p^{all} is the number of NU days.

Similarly, if q[th] hour is *Not Useful(NU),* then the.q^{th} column elements of matrix [X] matrix is called;

$$[X_{hour}^{NU}] = [x_{q,1} \quad \cdots \quad x_{q,n}]$$

If more than one day is *Not Useful(NU)*, then matrix $[X_{hour}^{NU}]$ is a $(q^{all} \times n)$ matrix where q^{all} is the number of NU hours.

Step 2: Obtain submatrix$[X^{reduced}]$ for the following case;

- Without *Not Useful* day data
  Remove matrix $[X_{day}^{NU}]$ from matrix [X] to form a submatrix $[X_{allhour,withoutNUday}^{reduced}]$ (note: *NU- Not Useful*)
- Without *Not Useful* hour data
  Remove matrix $[X_{hour}^{NU}]$ from matrix [X] to form a submatrix $[X_{withoutNUhour,allday}^{reduced}]$
- Without *Not Useful* day and *Not Useful* hour data
  Remove row matrix $[X_{hour}^{notuseful}]$ from matrix [X] to form a submatrix $[X_{withoutNUhour,withoutNUday}^{reduced}]$

Step 3: Calculate the mean and standard deviation of the electricity consumption for m days for matrix [X] and submatrix$[X^{reduced}]$ of each case and fill in the Elements of Data Selection Test (DST) Matrix by using the equation given in Table 1.

Table 1 Elements of Data Selection Test (DST) matrix

$DST_{r,s}$	Do not Remove $[X_{day}^{NU}]$	Remove $[X_{day}^{NU}]$
Do not Remove $[X_{hour}^{NU}]$	$\frac{\sigma}{\bar{y}}$	$\frac{\sigma_{allhour,wihoutNUday}^{reduced}}{\bar{y}_{allhour,wihoutNUday}^{reduced}}$
Remove $[X_{hour}^{NU}]$	$\frac{\sigma_{wihoutNUhour,allhour}^{reduced}}{\bar{y}_{wihoutNUhour,allhour}^{reduced}}$	$\frac{\sigma_{wihoutNUhour,wihoutNUday}^{reduced}}{\bar{y}_{wihoutNUhour,wihoutNUday}^{reduced}}$

Step 4: Determine $min_{r,s}\{DST_{r,s}\}$.

The elements with the lowest value (min) will determine the best data selection for predicting electricity consumption of a building. For example, if element $DST_{2,2}$ (in Table 1) is the lowest, $[X_{day}^{NU}]$ and $[X_{hour}^{NU}]$ must be removed from the historical data set to get better prediction. On the other hand, if if element $DST_{1,1}$ is the lowest, all data must be used.

5.0 CASE STUDY

5.1 Electricity Consumption Data

To test the applicability of the presented Data Selection Test (DST) method, an actual load profile data of a commercial building is used. Hourly consumption data of a university building (Selected blocks, Faculty of Electrical Engineering, UTM) is recorded for 2 months; April 2013 and May 2013 and are presented in Figure 2 and Figure 3 respectively. The April data whichis given in Figure 2 is used as the historical data to predictthe daily workday consumption forMay 2013. Itcanbeseenthattheelectricityconsumptionofthebuildingvariesovertimewithacertainpattern. It is stressed in this paper that the historical data (ie. April data) must be firstly tested to identify the correct data set before it can be used by any prediction method. ThepresentedData Selection Test (DST) method will be usedtoselectthedataset,thenusingit to predict the building electricity consumption for May 2013. Threestatistical prediction methods(ie. Moving

Average, MA, Exponential Smoothing, ES and Linear Regression, LR) is used. The predicted results are then compared to the actual consumptiondataof month May 2015.

5.2 Data Selection Test (DST) Matrix

It is assumed that Saturday andSunday (weekend) data is *Not Useful (NU)*. Also, it s assumed that hours outside working hours data (6pm-7am) are*Not Useful (NU)*. The April 2013 data is tested by using the presented DST method. BuusingStep 1-step 4 ofsection 4, the calculated each element of the DST Matrix is given as follows:

$$DST_{i,j} = \begin{bmatrix} 56.52\% & 10.48\% \\ 63.29\% & 9.87\% \end{bmatrix}$$

The DST matrix shows that element $DST_{2,2}$ gives the lowestpercentagevalue. Thisindicates that Saturday andSunday (weekend) data and hours outside working hours (6pm-7am) data must be removed from the historical data set. Thenewdataset, Data Set: {Remove Outside Working Hours Data AND Remove Weekend Data} must be used for prediction for better accuracy.

5.3 Prediction Accuracy Results by Using Different Data Set

To test the applicability of the DST method, the data set suggested by DST, as well as other data sets are used to predict the electricity consumptionofthebuildingformonth: May 2013 Three statistical prediction methods ie. ES, MA & LR presentedin sub-section 3.1 are used on the data sets. Statistical test; RMSE, MAPE, MAD and MRE are used to assess the prediction accuracy. The accuracy of the prediction for the three methods for different data set is presented in Table 2. The results show that all prediction methods give lowest error when using the data set suggested by the DST method (ie. Remove outside working hours data and remove weekend data).

The DST matrix result in subsection 5.2 also shows that element $DST_{1,2}$ gives the 2nd lowest value. This matrix element represents Data Set= {Remove Weekend Data Only}. The result in Table 2 shows that the error resulted by using this data set is second lowest. The results proved that the DST method is applicable in selecting the correct data for predicting electricity consumption of a building.

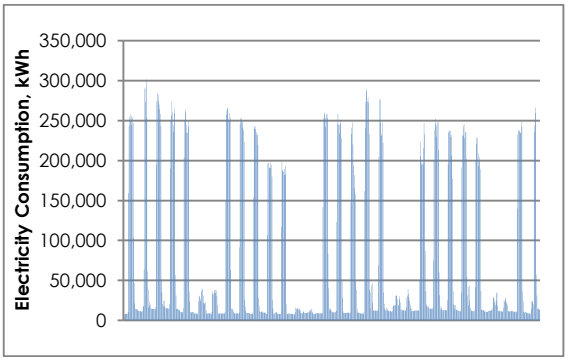


Figure 2 Actual hourly electricity consumption (kWh) for April 2013

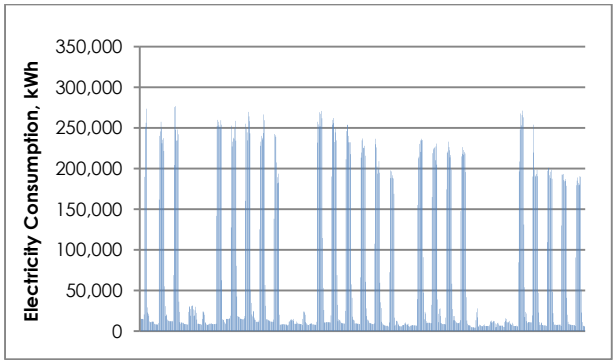


Figure 3 Actual hourly electricity consumption (kWh) for May 2013

6.0 CONCLUSION

A Data Selection Test (DST) has been presented in this paper. The method can be used to analyse the historical electricity consumption data to select the correct data set for prediction. The method has been tested on actual electricity consumption data. It is proven that the data set suggested by the DST method can improve prediction accuracy

Table 2 Prediction Accuracy by using various methods by using different data set.

Data set	Prediction method	RMSE	MAPE	MAD	MRE
All data	ES	2.94	1.66	8.00E+06	8.63
	MA	0.34	0.81	5.36E+06	3.21
	LR	2.02	1.31	1.15E+07	4.09
Remove weekend	ES	0.08	0.07	1.98E+06	0.01
	MA	0.02	0.09	2.58E+06	0.01
	LR	0.19	0.16	4.39E+06	0.04
Remove outside working hours	ES	5.67	3.09	7.06E+06	32.17
	MA	3.35	1.36	4.88E+06	11.23
	LR	4.00	2.35	1.07E+07	16.01
Remove outside working hours and weekend	ES	0.08	0.07	1.59E+06	0.01
	MA	0.10	0.08	1.88E+06	0.01
	LR	0.17	0.14	3.40E+06	0.03

Acknowledgement

The authors would like to thank for the support given to this research by Malaysian Ministry of Energy, Green Technology and Water (KeTTHA) and UniversitiTeknologi Malaysia (UTM), vote no. 4B111.

References

[1] Tenaga, S., 2012. National Energy Balance. *Suruhanjaya Tenaga (Energy Commission): Putrajaya, Malaysia*.
[2] Ben-Nakhi AF, Mahmoud MA, 2004. Cooling Load Prediction For Building Using General Regression Neural Networks. *Energy Conversion and Management*. 45(13-14): 2127-41.
[3] Wong SL, Wan KKW, Lam TNT, 2010. Artificial Neural Networks For Energy Analysis Of Office Building With Daylighting. *Applied Energy*. 87(2): 551-7.
[4] Aydinalp M, Urgusal VI, Fung AS, 2002. Modeling Of The Appliance, Lighting, And Space Cooling Energy Consumption In The Residential Sector Using Neural Networks. *Applied Energy*. 71(2): 87-110.
[5] Languang Z., Jing H., Jinxiang P., and Fengzhong Z., 2012. Electrical Energy Demand Forecasting With GRNN For Energy Saving Strategy, *Applied Mechanics and Materials* 198-199: 639-643.
[6] Victor M., Gareth A. Taylor, and Arthur E., 2014. A Novel Econometrics Model For Peak Demand Forecasting, *IEEE*.
[7] Kang J., Hong S., and Jian C., 2011. Forecasting Building Energy Consumption Using Neural Networks And Hybrid Neuro-Fuzzy System: A Comparative Study, *Energy and Buildings* 43: 2893-2899.
[8] Ahmad S. A., Muhammad Y. H., and Md. Shah M., 2012. Application Of Hybrid GMDH And Least Square Support Vector Machine In Energy Consumption Forecasting, *IEEE, International Conference on Power and Energy (PECon)*.
[9] S. Mohammadi, H. Keivani, M. Bakhshi, A. Mohammadi, M. R. Askari, and F. Kavehnia, Demand Forecasting Using Time Series Modeling and ANFIS Estimator.
[10] Arindrajit Pal, JyotiPrakash Singh, ParamarthaDutta, 2013. The Path Length Prediction of MANET Using Moving Average Model, *Procedia Technology*. 10: 882-889.
[11] Everette S. Gardner Jr., 2006. Exponential Smoothing: The State Of The Art—Part II, International Journal of Forecasting, 22(4): 637-666.
[12] Shalabh. 2013. A Revisit To Efficient Forecasting In Linear Regression Models, *Journal of Multivariate Analysis*. 114: 161-170.

.