



Forecasting Model of the Electricity Consumption: Preliminary Study using Statistical Data Analysis and Artificial Neural Network (SDA-ANN) Framework

Muhammad Asraf H.¹, Nur Dalila K.A.^{1*}, Nor Kamariah Kasmin@Bajuri², Zakiah M.Y.¹, Siti Mariam Mohammad Ilyas³, Nooritawati M.T.⁴

¹Faculty of Electrical Engineering, Universiti Teknologi MARA, 81750 Masai, Johor, Malaysia

²Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 81750 Masai, Johor, Malaysia

³Academy of Language Studies, Universiti Teknologi MARA, 81750 Masai, Johor, Malaysia

⁴Faculty of Electrical Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

*Corresponding author E-mail: nurdalila306@johor.uitm.edu.my

Abstract

There is a growing trend in utilizing forecast model in predicting energy usage in order to pursue sustainable practice. Among the benefit of the prediction model includes energy efficient usage and basic reduction in operating cost. The objective of this paper to employ Artificial Neural Network (ANN) to be used as the preliminary forecast model for electricity consumption trend prediction. The proposed ANN model will also include on pre-processing of the input variable through Statistical Data Analysis method to improve on the model. The model is then subjected to case study of electricity consumption in Universiti Teknologi MARA Pasir Gudang (UiTMPG), where the environmental temperature, number of student and staff noted as attendance data is designated as the input variables. Actual monthly billing data and UiTM attendance data is used as the testing data for the ANN, and the forecast model successfully generates the prediction equation. Results from the forecast model show the recorded Mean Relative Error (MRE) is within the acceptable value of 3.0254%.

Keywords: Artificial Neural Network (ANN); electricity consumption; prediction model; statistical data analysis (SDA); forecast

1. Introduction

An aspect of sustainable practice, which has been made crucial for operating companies or the government sectors in light of the global warming, includes good energy consumption management. A basic approach may include removing energy waste, either by decreasing electricity usage, or to efficiently use energy source by matching the peak electricity supply to production demand. Therefore, there is a need to be able to predict future utilisation with the given current operating trends, which may incur some costs in research and development. Fortunately, reduction in energy utilisation [1, 2] greatly lowers the operating costs of machineries and equipment, making it a worthwhile effort to formulate a reasonably accurate forecasting model for electricity consumptions. This then shall be the basis for the related management parties to make decisions regarding operations involving high cost value, or to readjust operation schedule to meet their production demand efficiently.

Previous study related to forecasting energy consumption has been initiated for decades. Various approaches have been developed in order to seek the most efficient forecast methods. Traditional forecasting technique had employed few methodologies such as time series, regression and soft computing techniques similar to neural networks[3-5]. Several other prediction attempts have applied non-linear time series, prediction model, classification, etc. The research problem to find the suitable methodology for prediction model remains ongoing. In this preliminary work, the

model of forecasting focuses in Artificial Neural Network (ANN) has been selected due to simple variables involves and does not involve high computational cost. As reported in [6,7] ANN has proven robust as compared to the complex computational algorithm. In addition, ANN possesses the ability to generalize nonlinear data well[8].

Analysis of electricity usage was carried out by A.Azadeh [9] by proposing a simulated ANN in their work. The data that randomly generated was compared with actual data which were further analyzed using ANOVA technique. The outcome proved that the superiority of ANN prediction as compared to time series techniques. Meanwhile another similar study related to using forecasting program to predict electricity usage have been supported by G.Tamizharasi [5] using the basic of ANN with different input. It was reported that long term energy consumption is able to be predicted based on four input variables - yearly ambient temperature, installed power capacity, yearly per resident electricity consumption and gross domestic product. The paper reports accuracy of 2% with the ANN method. Another research work to forecast electricity consumption was conducted by K.P. Amber[6], which compared several technique to forecast consumption over five year periods by using input parameters such as humidity, solar radiation, wind speed, temperature and weekday index. The result supports the use of ANN as a good model to forecast electricity consumption.

The main objective of this work is to design the forecasting model based on ANN to predict the next month of the electricity consumption. However, the scope of work for this paper will be

focusing merely within Artificial Intelligence (ANN) model integrated with Statistical Data Analysis (SDA). The improvement on the forecasting model comes from employing the Statistical Data Analysis (SDA) to reduce the unnecessary data through standard regression and correlation tests prior to feeding into the ANN environment. This methodology of processing the data using statistical analysis can be considered as a novel/innovative way of pre-processing the ANN input, and it is expected to be easier for the ANN forecasting model to perform analysis on any given data. Generally, the benefits of ANN are the ability of performing self-learning without pre-programming with a specific rule, which is loosely based on the biology of the human brain network which learns from the inputs from surrounding. To take advantage of imitating the brain neurons by ANN, this forecast model for electricity consumption will require proper data (or variables) as inputs to be processed by the SDA and learned by the ANN model. For this preliminary study, the selected input variables will be limited into attendance data and the environmental temperature data based on investigation in Section 2.

The rest of the paper contains the following: Section 2 describes the methodology implemented in this study. Section 3 presents the outcome of the forecast model and discussion towards findings. Section 4 concludes the paper in overall.

2. Methodology

The relationship between electricity consumption with several variables will be investigated. The main objective for this data analysis is to identify the effect of dependencies of these variables towards the energy consumption. For this to work, the investigated input variables are limited to three elements namely the environmental temperature, number of staffs, and number of students. Meanwhile the energy consumption acts as the output variable which is constrained from March 2016 and March 2017. The description of three variables is addressed as followings:-

- i) Environment temperature
The environment temperature encompasses the area of Universiti Teknologi MARA Bandar Seri Alam, Masai. The acquisition of temperature data in degree Celsius ($^{\circ}\text{C}$) is obtained from the website at <https://www.accuweather.com>.
- ii) Number of staff
The total working staff on duty throughout the duration of investigation is about 328 persons. However, the energy consumption trends do not correlate directly with the total working staff on payroll, instead the number of staff present in the real time shall be considered. Since the intention is to simulate and predict based on real time energy consumption, the availability of energy users at real time should be used as an input variable for the model. This data will be extracted from the facility e-attendance system and averaged out on the monthly basis.
- iii) Number of students
The number of students involved in this work consists of five different clusters of departments namely faculty of Business Management and Transportation, faculty of Electrical Engineering, faculty of Mechanical Engineering, faculty of Chemical Engineering and faculty of Civil Engineering. Similar to the staff, the consideration should be made for students' availability on real-time basis. However, the lack of attendance record limits this, hence assumption will be made as a full attendance of the student during the allocated time.
- iv) Electricity consumption
The usage of electricity is measured via billing consumption retrieved from UITM treasurer office. The electricity usage is

measured within the whole infrastructure building, denominated in Malaysia Ringgit (MYR) currency.

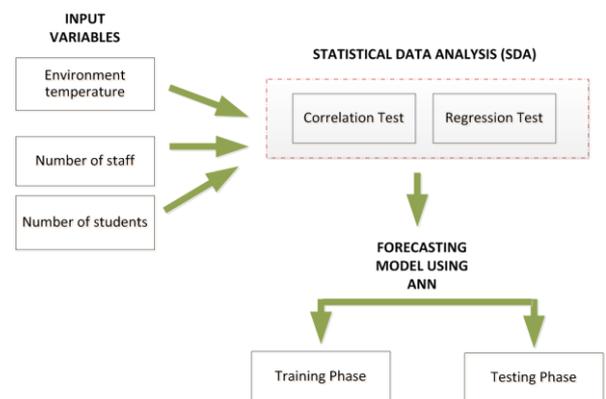


Fig. 1: Overall framework of the proposed system

An overall framework is summarised in Figure 1. The input variable will undergo pre-processing by SDA methods through what is called as correlation and regression test. These tests are to check whether there exists a relationship that shows input variable is affecting the output variable, as well as checking the strength of the relationship using linear regression method. Next, the prediction is performed by implementing ANN on the tested input variables.

2.1. Statistical Data Analysis (SDA)

Statistical analysis is used to identify which variables contribute to electricity usage. Several statistical models have been implemented by considering multivariable input as following:

$$\begin{aligned}
 Y &= \text{rate of electricity consumption (MYR)} \\
 X_1 &= \text{environmental temperature } (^{\circ}\text{C}) \\
 X_2 &= \text{number of staff (Pax)} \\
 X_3 &= \text{number of student (Pax)}
 \end{aligned}$$

Where Y is dependent variable and X is independent variable. Initial hypothesis related to objectives is stated as

$$\begin{aligned}
 H_0 &: \text{rate of electricity consumption, number of staff and} \\
 &\quad \text{number of student does not affect the electricity consumption} \\
 H_1 &: \text{rate of electricity consumption, number of staff and} \\
 &\quad \text{number of student affect the electricity consumption}
 \end{aligned}$$

One of the analysis uses correlation test to investigate the relationship between two variable which assigns a correlation value that is based on whether there exists a relationship between the input variables and the output. Table 1 defines the correlation value used in this analysis.

Table 1: Summary of correlation analysis and interpretation

Type	Correlation value	Definition
Negative correlation	-1 to 0	There is no relationship between two variables
Zero correlation	0	There is no relationship between two variables
Positive correlation	0 to 1	There exists the relationship between two variables

Regression test will be employed to determine the strength of relationship between the input and output. In this study, Multiple Linear Regression is selected to simulate electricity consumption trends for the forecasting model. The outcome of the test is to show the variability of the input with respect to the electricity consumption when multiple linear relationships is assumed. This

is represented via the summary output, Analysis of Variance (ANOVA) and the regression coefficient

[1] Summary Output

The elements exist in Summary Output are as follows:-

- a) **Multiple R** is a strength measure of the linear relationship between dependent variable and all independent variables. A stronger relationship exists in case of the value approaching to one.
- b) **R-square** is a measure in term of goodness of fit from regression equation. It represents the percentage of variation of dependent variable over independent variables. The value of R-square lies between zero to one and similar to Multiple R measure, the highest regression accuracy is shown when the R-square is one.
- a) **Adjusted R Square**. Function similarly to R-square measure. However, this measure is more accurate for the input which consists of more than one independent variable.
- b) **Standard Error**. Represent the discrepancy between predicted data and actual data. Lower value of Standard Error represents the more accurate of regression model.
- c) **Observations**. This demonstrates the number of data used in the work.

[2] ANOVA

The analysis of variance (ANOVA) works to examine the acceptability of regression model from the statistical perspective in the form of variance similarity [10]. The significance value is denoted as Significance F or Sig. F, which require that the value of Sig. F < 0.005 indicates that the independent variable is significantly effect the dependent variable. Therefore the ANOVA test will reject H_0 and accept H_1 .

[3] Regression Coefficient

The coefficient of regression shows the value of coefficients, standard error, t stat and P-value and time interval. These coefficients values will produce the regression model which later is used to predict data [11]. Note that the importance of the P-value in determining the significance of the dependent variable towards the independent variable. For P-value less than 0.05, the effect of dependent variable on the independent variable is significant.

2.2. Artificial Neural Network (ANN)

The artificial neural network consists of the neuron which is based on single perceptron model of the input neuron layer, output neuron layer and intermediate layer in between, called as hidden layer. The neurons in the model are mutually interconnected, each consists of respective weightage value. Initially, the input node will be multiplied with weightage to produce the sum value [12]. Subsequently the sum values will be passed to activation function to produce an output value. In this study, collected electricity consumption data will be used as training data and testing data for the ANN forecasting model. In training phase the interconnecting weights is adjusted until the error of the predicted model reached the most minimum value. Next, the error is propagated back to the network as a learning process. Therefore the ANN design will adopt a feedforward neural network with back-propagation learning algorithm [13]. The testing phase will be conducted after the training phase to evaluate the effectiveness of the developed forecasting model.

3. Results and Discussion

Analysis of data has been conducted to identify the effect of the variable namely environmental temperature and number of people towards electricity consumption in Universiti Teknologi MARA kampus Pasir Gudang (UiTMPG). Electricity consumption data collected within time period between March 16 until March 17 has been tabulated in Table 2. Regression and correlation tests have been conducted to investigate the relationship between variables. The test will determine whether to accept or reject the hypothesis of H_0 and H_1 . Correlation test provides the technique to measure the relationship strength between two variables. The measure indicates scale value from -1 to 1 which defines from no relationship to relationship exist between two variables as in Table 1. The outcome of that test can be referred in Table 3.

Table 2: Electricity consumption, temperature, student and staff attendance data collected at UiTMPG from March16 to March 17. These data shall be used as training data and testing data for the forecast model

Month	Environmental temperature (°C)	Number of staff (Pax)	Number of student (Pax)	Electricity consumption (MYR)
Mar 16	34	322	3226	226453
Apr 16	34	309	3222	259358
Mei 16	33	295	0	137989
Jun 16	32	318	3267	215630
Jul 16	32	322	3319	209580
Ogos 16	33	322	3323	251766
Sept 16	33	322	3311	208156
Oct 16	33	314	3311	213052
Nov 16	32	295	0	140131
Dec 16	32	322	3106	215401
Jan 17	32	322	3093	222259
Feb 17	31	322	3089	206801
Mar 17	32	322	3083	223751

Table 3: Correlation values for each input variables and among variables

Item	Environmental temperature (°C)	Number of staff (Pax)	Number of student (Pax)	Electricity consumption (MYR)
X_1	1			
X_2	-0.174	1		
X_3	0.0563	0.908	1	
Y	0.2844	0.734	0.894	1

Analysis that can be deduced from the Table 3 showed that the negative correlation value of -0.1735 evidently demonstrate there is no relationship exists between environmental temperature and number of staffs. Similarly, the correlation value between environmental temperature and number of students are low and close to zero (0.0563), indicating the absence of relationship between these two variables. However, between environmental temperature and electricity consumption, the correlation value is 0.2844, which points out links between the two, although it is considered as weak. Highest correlation can be observed between number of staff and students which is at 0.9081, which is explained by the reasoning that students and staff attendance trend tends to go hand in hand, being high during school season and drops together during semester break. From correlation testing, the results logically show that major contribution towards electricity consumption comes from the number of staff and students present respectively. Therefore, the environmental temperature can be excluded due to insignificance toward rising of electricity consumption. Table 4 shows the summary output from the regression test. The summary output shows that linear regression

is fairly reasonable approximation of relationship between the three variables (environment temperature, number of student/staff) with electricity consumption, since the multiple R value is calculated to be 0.9269. From the data, the best fit of the regression model is also conservative, with the R Square value of approximately 86% and adjusted R Square of 81.22%. Therefore 86% goodness of fit from regression model to proof the result is very good. Based on these parameters, the multiple linear regression model can be used to approximate the electricity consumption trend with respect to the environment temperature and the no of student/staff present.

Table 4: Summary output

Regression Statistics	
Multiple R	0.9269
R Square	0.8591
Adjusted R Square	0.8122
Standard Error	15293.1743
Observations	13

Table 5: Regression value to test the acceptance of the regression model

	df	SS	MS	F	Sig F
Regression	3	1.28x10 ¹⁰	4.28x10 ⁹	18.296	0.4 x 10 ⁻³
Residual	9	2.10x10 ⁹	0.24x10 ⁹		
Total	12	1.49x10 ¹⁰			

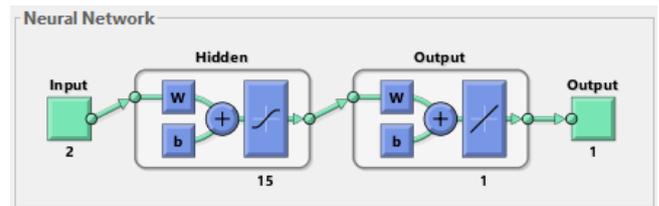
Based on Table 5, the Significance F value was found to be 0.004, which is in between 0 and 0.05. By this result, the H₀ is rejected and H₁ is accepted where all the input variables influence the electricity consumption in UiTMPG. Then, the regression coefficients were calculated and shown in Table 6 along with associated p-values, t Stat and standard error. The general equation for the forecast model is shown in equation 1.

$$Y = 96.48x10^3 + 7.64x10^3X_1 - 694.58X_2 + 31.02X_3 \quad (1)$$

Table 6: Coefficients value of each variables

	Coefficients	Standard Error	t Stat	P-value
Intercept	96.48x10 ³	0.5 x10 ⁶	0.1941	0.8504
X ₁	7.64x10 ³	5.98 x10 ³	1.278	0.2332
X ₂	-694.58	1.24 x10 ³	-0.5617	0.5880
X ₃	31.02	10.19	3.04	0.0139

Regression of data shows that for each one-unit increment of environmental temperature, the electricity consumption rises at RM7641.89. Next, the rate of consumption decrease is RM694.58 for each increase in staffs, while the contribution by student is RM31.02 for each increase in the student. Looking at the p-values for each variable, the most significant variable is the number of students, with the p-value of 0.0189, which satisfied the requirement of p-value less than 0.05. Other two variables failed to meet the p-value requirement. Further the forecasting model was tested using ANN technique for predicting the electricity consumption. A MATLAB software is implemented using NNTool box model and the following parameters as shown in Figure 2. Initially the normalization process has been conducted within the range of [-1,1] and then the activation function of transit function was applied. The next parameter namely input node, hidden node and output node was set at 2, 15 and 1 respectively. For the ANN, the data were divided as training and testing data for the network. The work considered randomly allocated 70% of data for training, 15% of data for validation and 15% for testing.



```

net      1x1 network
resultApril  2.1419e+05
t        1x13 double
testApril [322;3089]
tr       1x1 struct
x        2x13 double
    
```

```

net = fitnet(15);
view(net)
[net,tr] = train(net,x,t);
nntraintool
resultApril=net(testApril);
    
```

Fig. 2: Configuration of ANN used as forecasting model. The established model using NNTool box

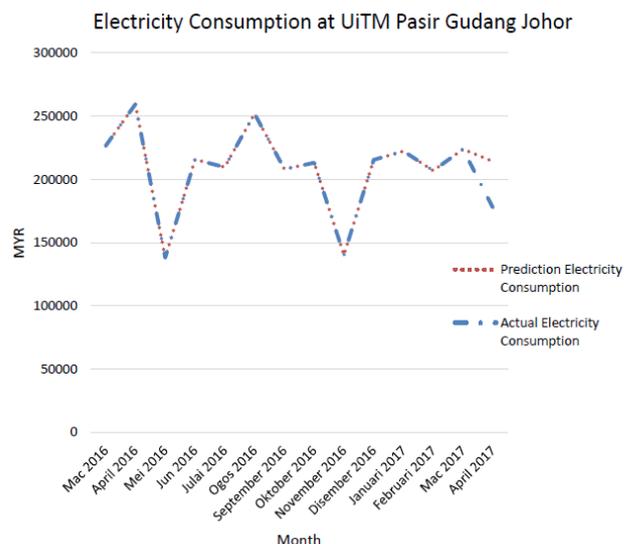


Fig. 3: Comparison of electricity consumption between actual electricity consumption and predicted values.

Comparison of predicted electricity consumption trend with actual billing is shown in Figure 3. The prediction model has successfully demonstrated the ability to estimate electricity consumption value based on given training data as input. In terms of performance forecast, the indicator of Mean Relative Error (MRE) demonstrated the value of 3.0254% which indicated low error within acceptable rate. The test on data has been conducted to forecast the bill at the month of April 2017, the forecast model predicts RM177593 of electricity usage, about 17% less than the actual electricity consumption billed.

4. Conclusion

This paper has proposed on employing Artificial Neural Network, with the input variables subjected to Statistical Data Analysis method to create a forecasting model with a given variable trend over time. The relationship between model was approximated using Multiple Linear Regression and a prediction equation was

successfully generated to predict the future trends from a given current data. From the case study of UiTM electricity consumption trend the predicted value using the model is observed to result a good estimation with MRE measure of 3.0254%. Future consideration for improvement may include to increase the possible input variable that affects the output, which require a separate intensive study.

Acknowledgement

The authors would like to thank Faculty of Electrical Engineering of Universiti Teknologi MARA (UiTM) of providing financial support for this work. Also the author acknowledges the Operational Excellence Team kampus Pasir Gudang whose name called Hamzah, Hj Basir and Syed Zahir and helpful team members for supporting the work.

References

- [1] F. Kaytez, M. C. Taplamacioglu, E. Cam, and F. Hardalac, "Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines," *International Journal of Electrical Power & Energy Systems*, vol. 67, pp. 431-438.
- [2] Y. T. Chae, R. Horeh, Y. Hwang, and Y. M. Lee, "Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings," *Energy and Buildings*, vol. 111, pp. 184-194.
- [3] L. Suganthi and A. A. Samuel, "Energy models for demand forecasting—A review," *Renewable and sustainable energy reviews*, vol. 16, pp. 1223-1240, 2012.
- [4] H.-Z. Li, S. Guo, C.-J. Li, and J.-Q. Sun, "A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm," *Knowledge-Based Systems*, vol. 37, pp. 378-387.
- [5] G. Tamizharasi, S. Kathiresan, and K. S. Sreenivasan, "Energy forecasting using artificial neural networks," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 3.
- [6] K. Amber, R. Ahmad, M. Aslam, A. Kousar, M. Usman, and M. Khan, "Intelligent Techniques for Forecasting Electricity Consumption of Buildings," *Energy*, 2018.
- [7] S. S. Baboo and I. K. Shereef, "An efficient weather forecasting system using artificial neural network," *International journal of environmental science and development*, vol. 1, p. 321, 2010.
- [8] P. P. Balestrassi, E. Popova, A. d. Paiva, and J. M. Lima, "Design of experiments on neural network's training for nonlinear time series forecasting," *Neurocomputing*, vol. 72, pp. 1160-1178, 2009.
- [9] A. Azadeh, S. Ghaderi, and S. Sohrabkhani, "A simulated-based neural network algorithm for forecasting electrical energy consumption in Iran," *Energy Policy*, vol. 36, pp. 2637-2644, 2008.
- [10] M. Sheikhan, M. Bejani, and D. Gharavian, "Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method," *Neural Computing and Applications*, vol. 23, pp. 215-227, 2013.
- [11] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses," *Behavior research methods*, vol. 41, pp. 1149-1160, 2009.
- [12] M. T. Sattari, K. Yurekli, and M. Pal, "Performance evaluation of artificial neural network approaches in forecasting reservoir inflow," *Applied Mathematical Modelling*, vol. 36, pp. 2649-2657, 2012.
- [13] N. Q. Hung, M. S. Babel, S. Weesakul, and N. K. Tripathi, "An artificial neural network model for rainfall forecasting in Bangkok, Thailand," *Hydrology and Earth System Sciences*, vol. 13, pp. 1413-1425, 2009.