Selangor Science &Technology Review
Special Issue: System Design & Instrumentation
Vol. 6, No. 3, (2022)

# Baseline Energy Model Development Using Artificial Neural Network: Small Dataset Approach

Farah Amira Zulfikri[1*], Wan Nazirah Wan Md Adnan[1]

[1]Department of Engineering, Faculty of Engineering and Life Sciences,
Universiti Selangor, 45600 Bestari Jaya, Selangor
wan_nazirah@unisel.edu.my

**Abstract**: The population in Malaysia is growing continuously, which is contributed to the increasing in electrical energy consumption. Due to that, there is important to reduce energy consumption and cost. Therefore, Measurement &Verification (M&V) baseline energy model in the baseline period is developed to calculate energy savings in the post-retrofit period for any energy management programs. Recently, Linear Regression (LR) is a method to develop the baseline energy model, but this method is less suitable for non-linear data. Artificial Neural Network (ANN) has been widely used in predicting and forecasting in various fields. This study is to develop baseline energy models for Option C International Performance of Measurement & Verification Protocol (IPMVP) using LR, ANN and ANN with resampling techniques to compare the performance of these three models based on their accuracy. The small dataset was chosen to examine the ability of the ANN model to train its network with less amount of data. Microsoft Excel was used to develop the LR model to correlate the energy consumption with several inputs which were working days, class days and Cooling Degree Days. MATLAB software was used to develop the ANN and ANN with resampling technique models using a single hidden layer with 3, 5 and 7 numbers of neurons in the hidden layer. The model with the highest accuracy were compared and analysed. Results show that ANN with resampling technique model is the best model to choose for a small dataset due to having the highest accuracy amongst the three models.

**Keywords**: Artificial Neural Network; Baseline Model; Energy Consumption; Resampling Technique

## 1. Introduction

Malaysia is one of the developing countries in South East Asia. The number of populations in Malaysia grew from 31,633,500 to 32,022,600 from year 2016 to 2017 (Mahidin, 2019). Due to that, the electricity consumption is also increased by 1.72 % in year 2017 as reported in Malaysia Energy Commission (Malaysia Energy Commission, 2017). Therefore, there is a need to reduce energy consumption while maintaining the productivity. This situation has encouraged Malaysian government to take several initiatives to promote and implement energy management programs to reduce energy consumption in Malaysia.

Measurement and Verification (M&V) is introduced to evaluate savings in energy management programs. M&V is the process of using measurement to reliably determine actual savings created within an individual facility. Savings cannot be directly measured, since they represent the absence of energy use. Instead, savings are determined by comparing energy consumption measured use before and after implementation of a project, making appropriate adjustments for changes in conditions.

A baseline energy is a reference tool that allows the energy saving companies to compare energy performance before and after retrofitting process. The baseline establishes the “before” by capturing a site or system's total energy consumption prior to making improvements. After the energy management programs implementation or during post retrofit period, this baseline energy is used to approximate how much the energy would have been used if there had been no energy management programs implementation.

There are four M&V options defined by the International Performance Measurement and Verification Protocol (IPMVP)(Efficiency Valuation Organization, 2012). Option A – Retrofit isolation with key parameter measurement, Option B – Retrofit isolation with all parameter measurement, Option C – whole facility, and Option D – calibrated simulation. It is important to develop the M&V baseline energy model in baseline period in order to accurately calculate energy savings for any energy management programs. This study emphases on development of baseline energy model for IPMVP Option C whole facility energy use and the data usually obtained from monthly utility bills.

Recently, Artificial Neural Network (ANN) has been widely used in predicting and forecasting in various fields. ANN is the part of a computing system where it is designed to simulate the way the brain processes information. This study proposes the development of baseline energy models using Linear Regression (LR) and ANN techniques. Comparison of both techniques will determine the most accurate baseline energy model.

## 2. Methodology

In this study, three baseline energy models, LR model, ANN model and ANN with resampling technique model were developed using Microsoft Excel and MATLAB software. A flowchart of the baseline energy models development is shown in Fig.1. Then, the results of these three models were compared and analysed based on their performance.

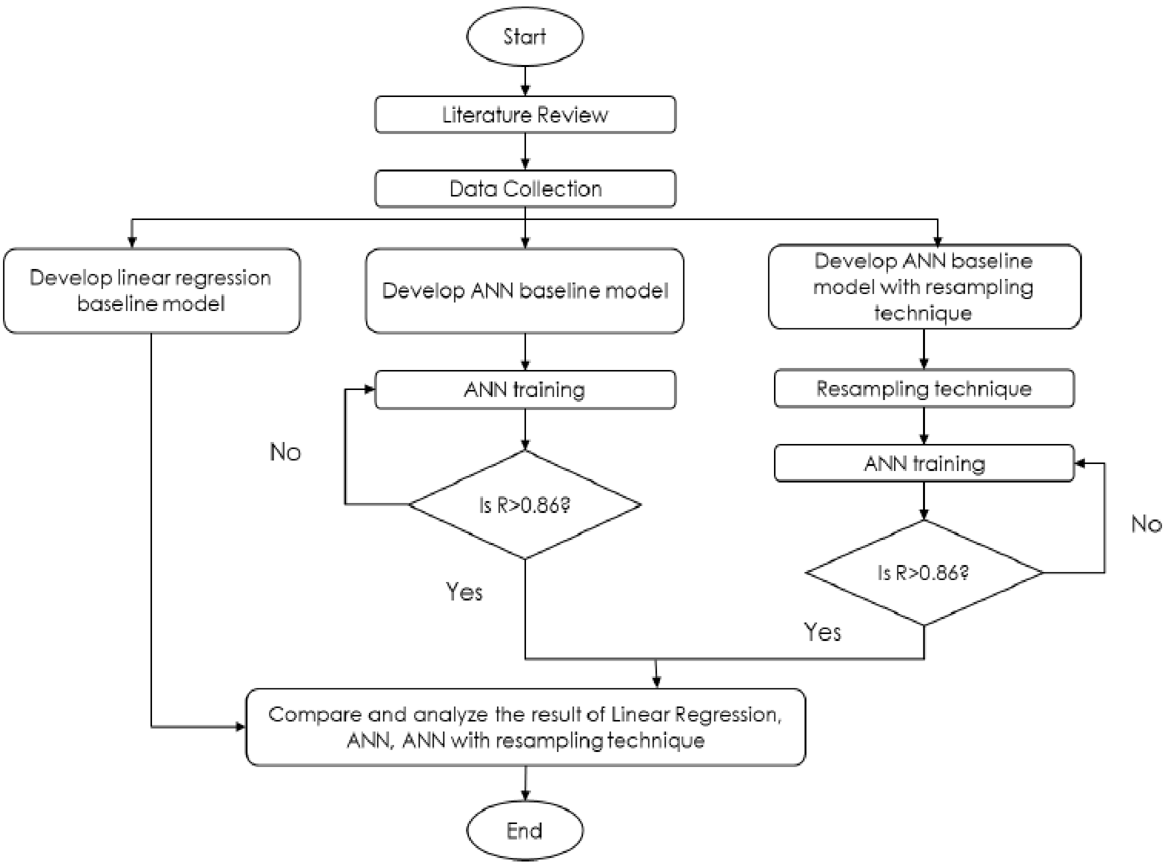


**Fig. 1** Flowchart of baseline energy model

### 2.1 Collection of Data

The data obtained from monthly utility bill of one of the university buildings in Selangor, Malaysia for 21 months. For the proposed model, three input variables that may influence the energy consumption were chosen, which are working days, class days, and Cooling Degree Days (CDD). Meanwhile, the output is energy consumption in kilowatt hour (kWh). In total, there were 21 data that consists of three different inputs and only one output.

## 2.2 LR Baseline Model Development

Microsoft Excel was used in order to develop the LR model to correlate the energy consumption as an output and with several input variable such as working days, class days and CDD. In Microsoft Excel, the data were assigned as input 1, input 2, input 3 and output. Input 1 was working days, input 2 was class days, and input 3 was CDD. Seven linear regression models were developed with a combination of several inputs and single output as presented in Fig. 2. The combination of inputs with the highest value of coefficient of determination ($R^2$) was chosen in order to develop ANN model as well as to compare the performance with the other models.

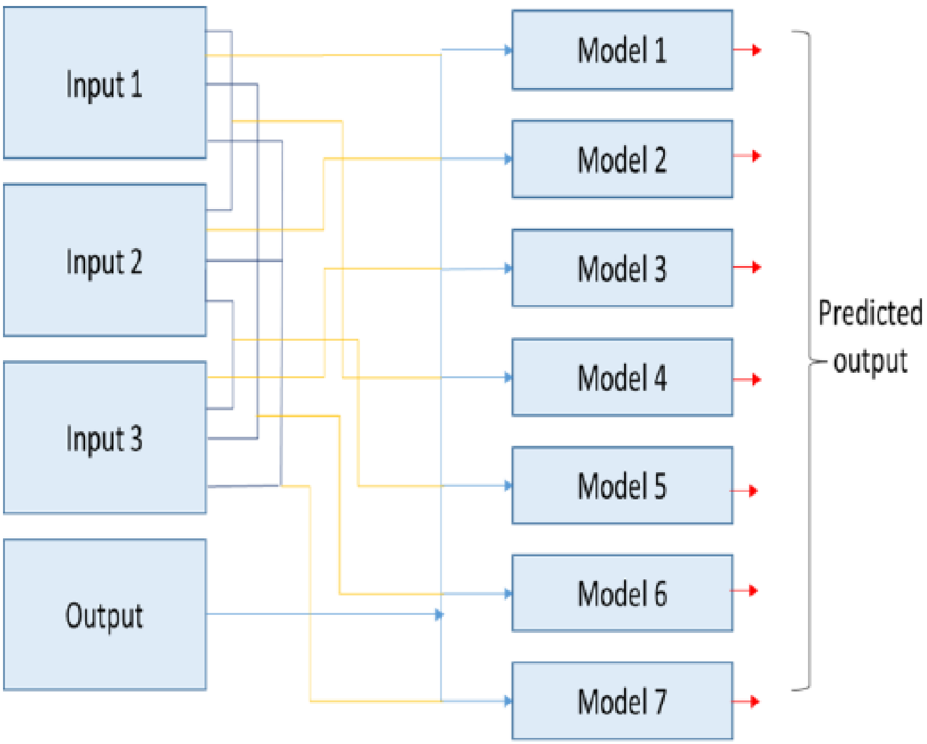


**Fig. 2** Regression model with different combination of input variables

## 2.3 ANN Baseline Model Development

ANN model was developed using MATLAB software using single hidden layer with 3, 5 and 7 number of neurons in the hidden layer. In the output layer the activation function used was sigmoid function and the ANN training algorithm was Levenberg-Marquardt (trainlm), where it is mostly used to train the network by several researchers (Gunawan et al., 2017; Tehlah et al., 2016). The data were separated into three parts, 70% for training process, 15% for validation process and another 15% for testing process (Adnan et al., 2019). After configuring all the ANN parameters, the next step was ANN training. The coefficient of correlation (R) was used as performance function to evaluate the model accuracy during ANN network training process. The equation for R is shown as in Eq. (1). Coefficient of Determination is the square of coefficient of correlation. For IPMVP, $R^2 > 0.75$ ($R > 0.86$) explains statistically significant performance (Newsham, 2019) and acceptable for baseline model development.

$$R = \frac{N(\Sigma y_t y_p) - (\Sigma y_t)(\sum y_p)}{\sqrt{[N(\sum y_t^2) - (\sum y_t)^2][N(\sum y_p^2) - (\sum y_t)^2]}} \tag{1}$$

## 2.4 ANN with Resampling Technique Baseline Model Development

In order to overcome the less accurate problem of ANN training due to the small dataset, resampling technique was added in ANN before the training process. There were few types of resampling technique but this study is only focuses on Cross Validation (CV) technique. The types of CV implemented in this study is K-Fold CV. In K-Fold CV the data were partitions into k-samples. Training and validation set were divided into k-subsample and in this study k=6 was selected. In this CV technique to train the data, first iteration, k-1 subsamples were devoted to train and validate the data and the remaining subsamples was used to test the date, and for k-iterations this procedure was repeated until all the data were used as test data (Ling et al., 2019). Similar to ANN baseline model, the number of neurons in the hidden layer was set to 3, 5 and 7 neurons with the same setting configurations. R was also used as to measure the model performance.

## 2.5 Performance Comparison

The performance of these three different models, LR, ANN and ANN with resampling technique were compared to evaluate the accuracy of the model. The comparison is based on the highest value of $R^2$ from LR and highest value of all Rs from ANN and ANN with resampling technique. The most accurate model was selected and used to determine the savings in post-retrofit period.

# 3. Results and Discussions

The purpose of this study was to develop, compare and determine the most accurate baseline energy model where this baseline energy model may use to calculate savings in post-retrofit period. The baseline energy models were developed using three different models, LR, ANN and ANN with resampling technique. Three input variables were considered, class days, working days and CDD where the output was energy consumption. The data were gathered from one of the universities in Malaysia.

The results for all models were compared as tabulated in Table 1. For LR model, the highest value of $R^2$ that this model can achieved was only $R^2$=0.77 (where R = 0.8825). By using ANN model in developing baseline energy, the combination of 7 number of neurons produced the highest value of R which was 0.92504 compare to 3 and 5 number of neurons in the hidden layer which were 0.91445, and 0.91259 respectively. For ANN with resampling using CV technique, the ANN-CV with 7 number of neurons obtained the highest accuracy, which was R=0.92677. It is found that the baseline energy model using ANN with resampling technique with combination of 7 neurons in hidden layer is the best baseline energy model with Rtrain of 0.94920, Rvalid of 0.82936, Rtest of 0.99599 and Rall of 0.92677. Fig. 3 shows the R plots of energy consumption for Rtrain, Rvalid, Rtest and Rall for the most accurate model obtained in this study which was ANN with resampling technique with the combination of 7 neurons in hidden layer.

**Table 1**. Performance Comparison of All Baseline Energy Models

| Model | Rtrain | Rvalid | Rtest | Rall |
|---|---|---|---|---|
| LR | - | - | - | 0.8825 |
| ANN – 3 Neuron | 0.93252 | 0.8803 | 0.23625 | 0.91445 |
| ANN – 5 Neuron | 0.99996 | 0.9676 | 0.87641 | 0.91259 |
| ANN – 7 Neuron | 0.98846 | 0.91745 | 0.95522 | 0.92504 |
| ANN-CV 3 Neuron | 0.86509 | 0.99079 | 0.99549 | 0.89851 |
| ANN-CV 5 Neuron | 0.89439 | 0.99635 | 0.99412 | 0.92105 |
| ANN-CV 7 Neuron | 0.94920 | 0.82936 | 0.99599 | 0.92677 |

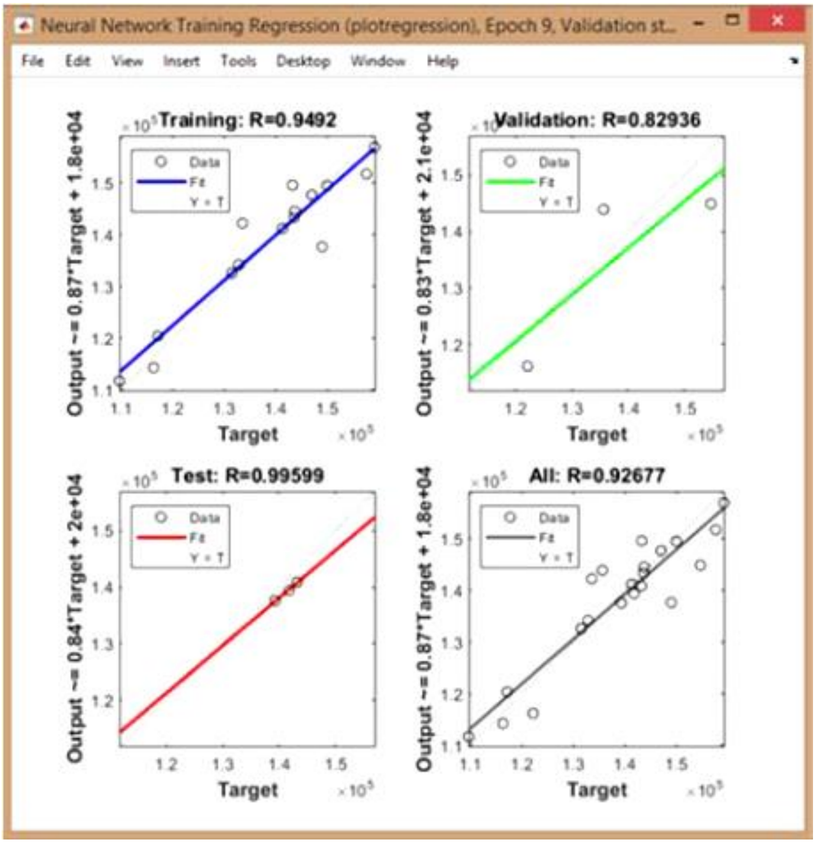


**Fig. 3** Rtrain, Rvalid, Rtest and Rall for ANN-CV 7 neurons

## 4. Conclusion

This study aims to improve the baseline energy model using ANN. The problem of inaccuracy in LR and ANN can be overcome by using ANN with resampling technique. The resampling technique CV was embedded to the ANN model in order to cope with inaccuracy when the dataset is small, especially for non-linear data. It is found that the proposed techniques to integrate ANN with resampling techniques give better-quality of prediction performance. Based on the results, ANN with resampling technique model with 7 neurons in hidden layer structure had significant accuracy improvements based on the highest values of R when observed in comparison to the other models. This model was selected as the baseline energy model to predict energy consumption for Option C IPMVP. The CV method was able to avoid overfitting, creating model diversity with limited dataset, and training neural networks based on the outcome of this study.

## 5. Acknowledgements

Special thanks to the department of Engineering, FELS, Universiti Selangor for the funding and the opportunity of presenting my research outcome in the FELS-US 2.0 2021.

## 6. References

Adnan, W. N. W. M., Dahlan, N. Y., & Musirin, I. (2019). Comparative Assessment of Artificial Neural Network Based Baseline Energy Model to Quantify Energy Savings of Chiller System in Commercial Building. *International Conference on Environmental Sustainability and Resource Security (IC-ENSURES)*, 58–62.

Efficiency Valuation Organization. (2012). *International Performance Measurement and Verification Protocol (IPMVP)* (Vol. 1).

Gunawan, T. S., Yaacob, I. Z., & Kartiwi, M. (2017). Artificial Neural Network Based Fast Edge Detection Algorithm for MRI Medical Images. *Indonesian Journal of Electrical Engineering and Computer Science*, *7*(1), 123–130. https://doi.org/10.11591/ijeecs.v7.i1.pp123-130

Ling, H., Qian, C., Kang, W., Liang, C., & Chen, H. (2019). Combination of Support Vector Machine and K-Fold cross validation to predict compressive strength of concrete in marine environment. *Construction and Building Materials*, *206*, 355–363. https://doi.org/10.1016/j.conbuildmat.2019.02.071

Mahidin, M. U. (2019). Department of Statistics Malaysia Press Release: Current Population Estimates, Malaysia, 2018-2019. In *Department of Statistics Malaysia* (Issue July).

Malaysia Energy Commission. (2017). *National Energy Balance 2017*.

Newsham, G. R. (2019). Measurement and verification of energy conservation measures using whole- building electricity data from four identical office towers. *Applied Energy*, *255*(September), 113882. https://doi.org/10.1016/j.apenergy.2019.113882

Tehlah, N., Kaewpradit, P., & Mujtaba, I. M. (2016). Artificial neural network based modeling and optimization of refined palm oil process. *Neurocomputing*, *216*, 489–501. https://doi.org/10.1016/j.neucom.2016.07.050