J. Sci. Univ. Kelaniya ***14*** *(2021): 17-30*

DOMESTIC ELECTRICITY USAGE ESTIMATION MODEL USING SOCIO-ECONOMIC FACTORS FOR SRI LANKA

Y.S.S. ARIYARATHNE[1], N.W.K. JAYATISSA[1] AND D.S.M. DE SILVA[2]

[1]Department of Physics and Electronics, University of Kelaniya, Sri Lanka

[2]Department of Chemistry, University of Kelaniya, Sri Lanka

Received 11th February 2021 / Accepted 26th May 2021

ABSTRACT

In this empirical study, socioeconomic factors that can easily be extracted from families have been used to build a "home electricity usage prediction" model based on two variables, family monthly income and family size. Each of these factors was evaluated individually. Two machine learning models were built using those factors as features. Two new parameters were introduced to analyze the data. Models are based on “Linear Regression” and “Random Forest” algorithms. This study revealed that the socioeconomic factors such as family size and family income are very effective in domestic electricity usage prediction model building, where the end usages are not known. Furthermore, the random forest algorithm was found to be more effective for unseen data than the linear regression algorithm. The accuracy of the models can be further improved by adding more data into the both models.

Keywords: Domestic electricity consumption, Electricity prediction model, Linear regression, Random forest, Machine learning

*Corresponding author Email: jayatissa@kln.ac.lk;

http://orcid.org/0000-0002-4515-1086

DOI: http:://doi.org/10.4038/josuk.v14i0.8031

INTRODUCTION

Electricity is the backbone of modern economies. Demand for electrical energy is rapidly increasing in the developing world. An electrical system consists of generating, transmitting, and distributing electrical energy. This process remains complicated and costly, so meeting its increasing demand has become a significant challenge to every nation in the modern world.

Electricity users can be grouped into several sectors. Therein, the residential demand constitutes a large percentage of the overall electrical energy demand. For example, domestic buildings consume around 40% of global generated energy (Saberbari & Saboori, 2014). Domestic electricity demand is one of the essential variables required for estimating the amount of additional capacity required to ensure a sufficient supply of energy. The right grid management strategies should involve load demand planning and an appropriate schedule for generating an effective load distribution. Accurate electricity demand forecasting should be used to maximize the efficiency of the planning and strategy formulation process in the power of domestic distribution systems (Nti, Teimeh, Nyarko-Boateng, & Adekoya, 2020). Therefore, load prediction and forecasting have become one of the major research fields in electrical engineering.

Many computational and statistical techniques have been applied to enhance forecast models to improve the estimations' accuracy and reliability. Numerous studies on estimating electrical energy demand for residential and commercial purposes have been conducted to enable electricity generators, distributors, and suppliers to plan effectively ahead and promote energy conservation among the users (Nti, Teimeh, Nyarko-Boateng, & Adekoya, 2020; Saberbari & Saboori, 2014). Two main types of load estimation approaches have been developed, known as correlation methods and extrapolation methods (Nti, Teimeh, Nyarko-Boateng, & Adekoya, 2020).

The extrapolation or trend analysis techniques involve fitting the trend curves to electricity demand time series data, and the future value of demand is obtained from estimating the curve function at the preferred future point. These approaches can be named as forecasting approaches (Aleksandar, Sonja, & Ljupco, 2016; Nti, Samuel, Michael, & Asafo-Adjei, 2019; Nti, Teimeh, Nyarko-Boateng, & Adekoya, 2020).

Correlation methods relate the electricity demand to several social, economic, and other demographic factors. These methods ensure that the analysts capture the relationship existing between demand variation patterns and other measurable factors. Predicting electricity demand using social, economic, and demographic factors are more complicated than the load forecast using time series data (Aleksandar, Sonja, & Ljupco, 2016; Nti, Samuel, Michael, & Asafo-Adjei, 2019; Nti, Teimeh, Nyarko-Boateng, & Adekoya, 2020). Because many factors can affect electricity demand, in these prediction methods, usually social-economic factors such as population, building sizes, employment data, weather data, building structure, and business types are used as the variables that define the electricity demand (Çamurdan & Ganiz, 2017).

Residential electricity energy is mainly used for heating, cooling, lighting, cooking, and entertainment purposes. Variations in electricity demand in tropical countries such as Sri Lanka are primarily influenced by socioeconomic factors and not by the vagaries of nature. Because in tropical countries close to the equator, temperature changes are within only several degrees range. Moreover, the seasonal changes of weather will not affect lifestyles that much. That can be easily identified by analysing the load curves over the year. As an example, changes in daily load curves in Sri Lanka are almost the same for every month (Amarawickrama & Lester, 2007).

The purpose of this research is to examine the effectiveness of socioeconomic factors in predicting domestic electricity demand. Family income and family size were selected as the factors for determining the electricity demand of a family. These two factors were chosen because the findings of several energy demand studies show that income and household size have a significant influence on energy consumption levels (Rajmohan & Weerahewa, 2010), and the data collected also shows a greater relationship between power usage and family income and power usage vs family size. Rather than the above facts it is always easy to collect that kind of simple data such as the family size and family income in many practical situations without much effort. So the prediction model based only on those two factors will be beneficial in many practical situations, where the end usages and other relative factors are unknown. In the first part of this paper, each factor is individually fit to the ordinary least squares (OLS) linear regression model and effectiveness of each factor as a prediction variable is individually analysed. Then both factors are combined and a multiple ordinary least square linear regression model is developed. Then the Random forest algorithm has been used for modelling residential electricity consumption, and finally, the effectiveness of the model is compared.

METHODOLOGY

Data Collection

The first part of this research is the data collection. There are about 51 million of households in Sri Lanka (Department of cenus and statics, 2016). Two methods are used to collect the data about the electricity usage of those households. As the first approach, printed questionnaire was distributed. This includes the questions about the family details and electricity usage of the home. And as the second approach an online form was shared. Using both of these methods 820 data sets were collected by covering all 25 districts in Sri Lanka. Data is collected from pure domestic uses, who do not use household electricity for any commercial purpose and only depend on main grid power.

Collecting data about individual household income is a somewhat challenging task. The average monthly income of Sri Lankan families was LKR 62,237 per month (Department of cenus and statics, 2016). This value includes wages and salaries and all the other sources such as agricultural and non-agricultural activities and all other monitory receipts.

In this study, the data were collected under four monthly income categories with LKR 50,000 range;

1. Below LKR 50 000,
2. LKR 50 001 – LKR 100 000,
3. LKR 100 001 – LKR 150 000,
4. LKR 150 001 – LKR 200 000.

Data Analysis

According to the "Household income and expenditure survey 2016", since the most recent "monthly family income" (MFI) is LKR 62,237, this can be related to income level range used in this research (IRZ) as follows.

$$\text{IRZ} = \frac{MFI}{1.25} \text{ (round to nearest 1000).} \tag{1}$$

The upper margin and Lower margin of each level n (UMn and LMn) can be calculated as follows.

$$\text{Upper margin of level n -: } \mathrm{UMn} = \mathrm{n(IRZ)} \tag{2}$$

$$\text{Lower margin of level n -: } \mathrm{LMn} = \mathrm{UMn} - \mathrm{LM1} + 1 \tag{3}$$

Electricity Demand Modelling

Prior to the analysis, the data set was pre-processed by removing the outliers. A new parameter (NP) was introduced to identify the outliers.

$$\mathrm{NP} = \frac{\text{Monthly Power Consumption in kWh}}{\text{Number of Family members} \times (\text{lower margin of the income catogary} + \text{MFI})} \tag{4}$$

To remove the anomalies, all the data away from mean to one standard deviation were removed based on the variable NP. In this step 188 data samples were removed, which was about 23% of the sample.

Family income.

The average monthly power consumptions for families in each income level are shown in Figure 1.

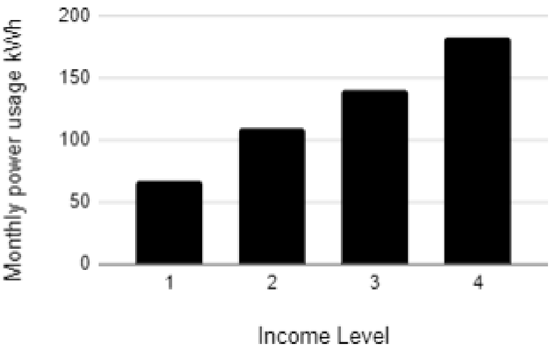

Figure 1: Average monthly electricity usage (kWh) of each income group.

Usually household energy consumption increases with the family income, so it is reasonable to use family income as a factor to predict the energy usage of family. Here the lowest income group uses a few electrical appliances and when the income goes high people tend to use more appliances. Hence more electricity to have more comfort in life.

To find the relationship between power usage and income, a new parameter (Income Parameter: IP) was introduced.

$$\mathrm{IP} = \frac{LMn+MFI}{MFI} \tag{5}$$

The Income Parameter (x) and power consumption (y) were fitted into an ordinary least squares regression model and equation (6) was derived, to analyse the relationship among the factors.

$$y = 45.4x + 21.8 \tag{6}$$

Here, r^2, coefficient of determination percentage of the dependent variable change, is equal to 0.51.

Family size

The other main factor used in this study is the number of family members or family size. Figure 2 shows how the power consumption increases with the number of family members in each house. Here, the family size and the monthly income of a family are also correlated.

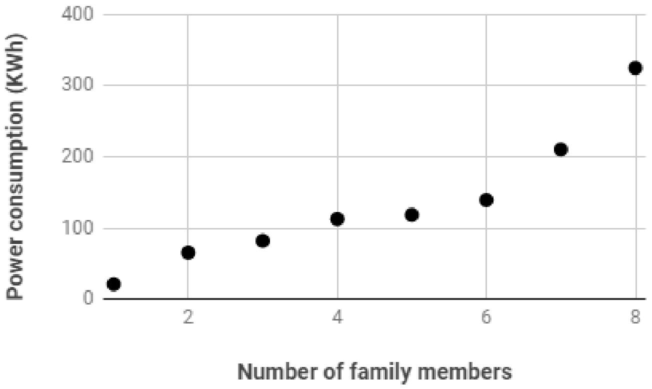

Figure 2: Number of family members vs. Average monthly power consumption.

This graph shows that power consumption gradually increases up to family size 6, and families that have more than six members show sudden increase. The family size (x) and power consumption (y) were fitted into an ordinary least squares regression model and relationship was derived.

$$y = 22.1x + 24.8 \tag{7}$$

Here, r^2, coefficient of determination percentage of the dependent variable change, is equal to 0.204 which is very low.

When comparing the above two models, it shows that least squares regression model does not correlate the income level and the power consumption much. As the data depends on socio economic factors, this kind of low value can be expected (Minitab, 2013).

Machine learning model building

Training the parameters of a prediction function and testing it on the same data is considered as a methodological mistake in machine learning: a model that would only repeat the labels of the samples that it has just seen may give a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called overfitting (Varoquaux, 2013). In machine learning model building a data set is usually divided into two parts; a training set and a testing set. Training data is used to build the machine learning model, whereas testing data is to evaluate the model performance. Here 80% of data were used to train each algorithm, and the rest were used to evaluate the model performance.

To avoid overfitting furthermore, the "Shuffle split" cross-validation method was also used. The "Shuffle split" iterator will generate a user-defined number of independent train/test datasets. Data are first shuffled and then split into a pair of train and test sets. This allows finer control on the number of iterations and the proportion of samples on each side of the train/test split. In this study, the data set was shuffled and split into train and test sets randomly for five times. The shapes of each train and test set are shown in Figure 3.

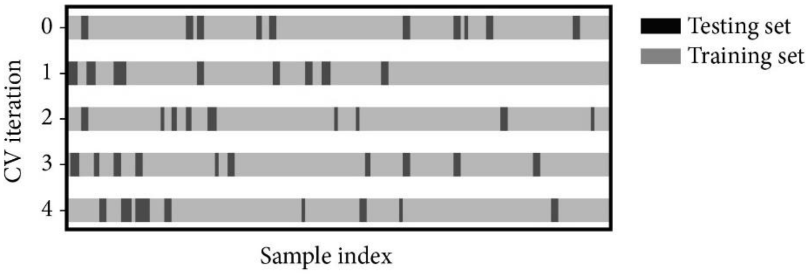

Figure 3: Shape of each train and test iteration created by shuffle split method.

Since this is a supervised machine learning model, the algorithm was given both the features (family size and income) and the targets (power consumption). During training, the features and targets are fed into two algorithms: linear regression and random forest. Both algorithms are trained to map correlations between features and targets.

RESULTS AND DISCUSSION

Linear regression model

The pre-processed data has been fitted into a multiple ordinary least squares regression model and derived relationship (4) among the Income Parameter (x), Family size (z) and power consumption (y).

$$y = 45.61x + 22.32z - 73.51 \qquad (8)$$

The r^2 of this model is 0.72 for the test data. That means when the power consumption is changed by a unit, about 72% of this change can be predicted by using the two variables, family size and monthly income level.

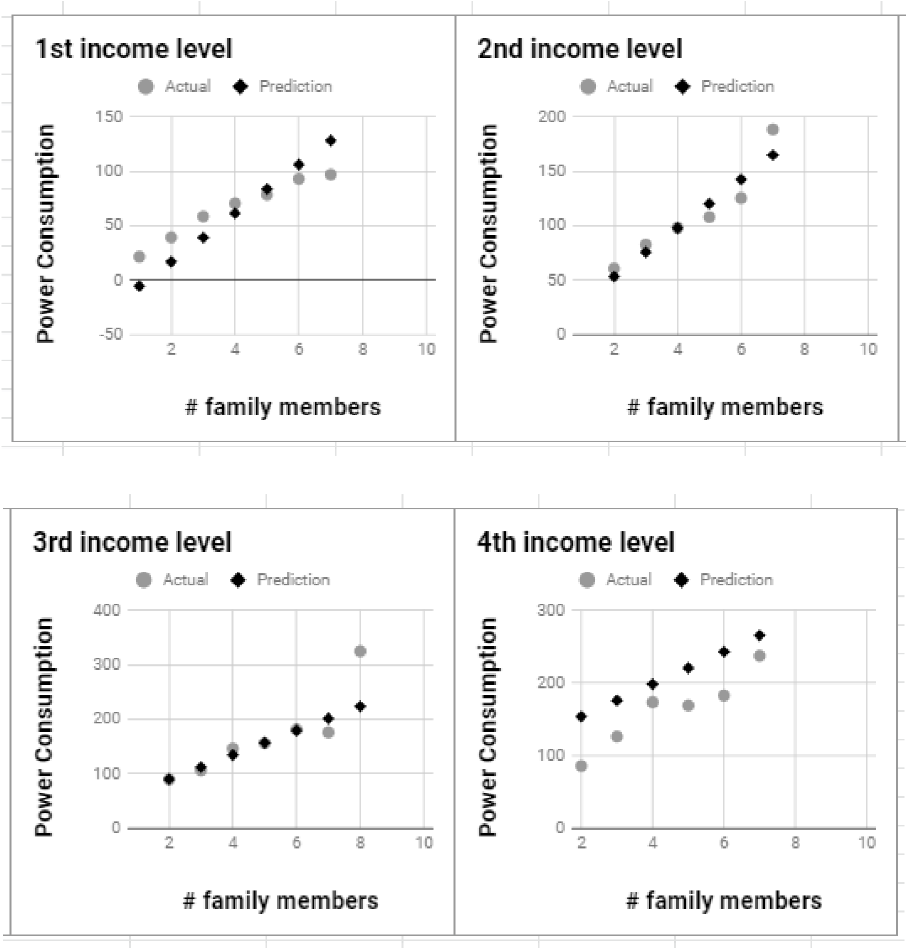

Figure 4: Average actual power usage (both train and test data) and predicted power usage of the for each income level using OLS regression model

The "Pearson product-moment correlation coefficients" between the mean of the actual values and predicted value for family monthly electricity usage of each income level are as follows.

- Level 1: 0.98
- Level 2: 0.94
- Level 3: 0.88
- Level 4: 0.95

As an average, this linear model has 0.94 Pearson product-moment correlation coefficients. The r^2, is the square of the Pearson correlation coefficient (r). But as mentioned above, for the same test data, linear regression model only gives 0.72 as the value for r^2.

Random forest algorithm

The idea of the Random Forest algorithm was introduced by Tin Kam Ho in 1995 (Ho, 1995). The decision tree is the building block of a random forest algorithm, and it is an intuitive model. Simply decision tree is a series of yes/no questions asked about the data, eventually leading to a predicted value. The Decision Tree algorithm has a significant disadvantage in that it causes over-fitting. This problem can be limited by implementing the Random Forest Regression in place of the Decision Tree Regression. Random Forest works by combining many decision trees in training data so that it will produce a high level of accuracy. Random forest is an interpretable model because it makes predictions much as humans do, like asking a sequence of queries about the available data until reaching a decision in an ideal world (Koehrsen, 2018).

Predictions of the random forest model and the mean of actual value of each income group is shown in Figure 5.

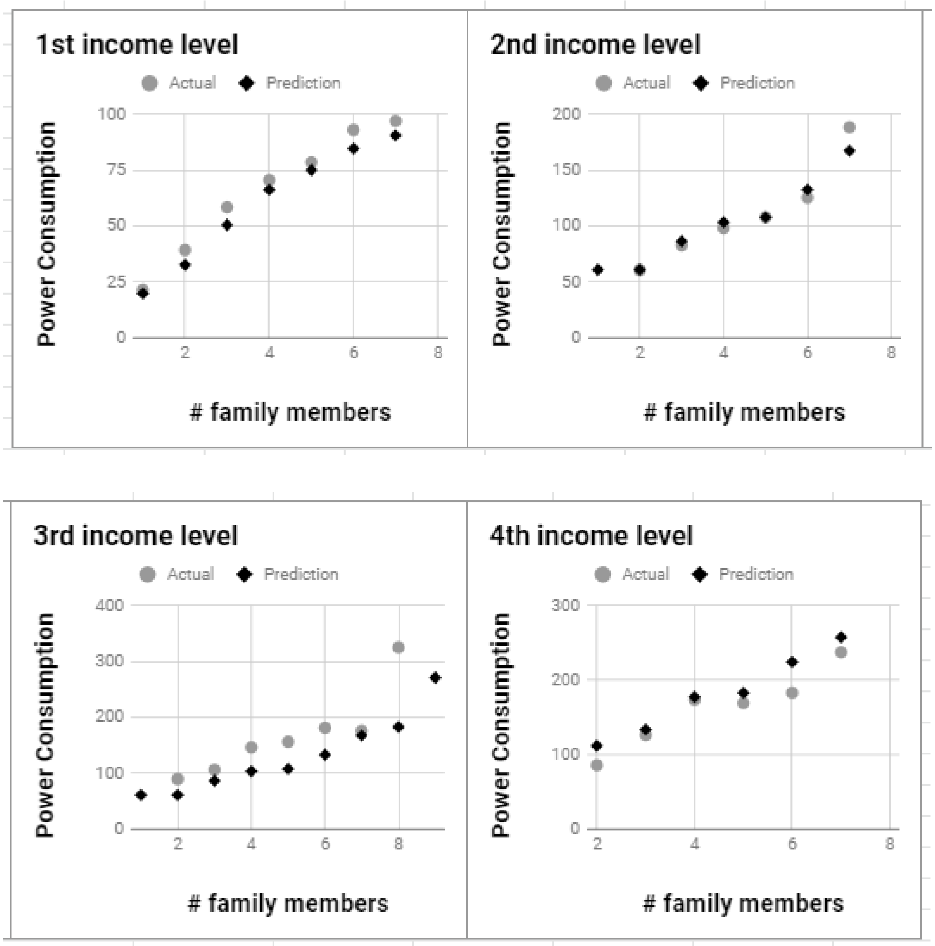

Figure 5: Average actual power usage and predicted power usage of each income level by using random forest algorithm.

The prediction, most of the time lies under the mean of the actual values. The "Pearson product-moment correlation coefficients" between the mean of the actual values and the predicted values for family monthly electricity usage for different income levels, shown in Figure 5, are as follows.

- Level 1: 0.99
- Level 2: 0.98
- Level 3: 0.87
- Level 4: 0.96

As an average, this linear model has 0.95 Pearson product-moment correlation coefficients. For the given training data, the "Random Forest" algorithm can make the predictions of electricity usage with an r^2 value of 0.76 by only taking family income level and number of family members as inputs. That means when the power consumption is changed by a unit, about 76% of this change can be predicted by using the two variables, family size and the monthly income.

Table 1: Comparison of algorithms.

Scoring method	Linear Regression	Random forest regression
R squared (between mean of the actual values and predicted values)	0.88	0.90
R squared (between the test data and predicted value)	0.72	0.76

By comparing the above results, it's clear that both regressions models have high success rates for the predictions. However, the random forest algorithm shows higher scores than the linear regression. These model accuracies must be further improved by adding more data points to the training set. However, as an initiative, the above results are acceptable.

CONCLUSIONS

By analysing the score for each algorithm it is clear that a successful domestic power usage prediction model can be built using the socio-economic factors, family size and monthly family income. At the bigger family sizes (between 8-10 members), the prediction seems to have a higher deviation than that of small family sizes. That maybe due to the insufficient data values to train the model in that range as, the average Sri Lankan family size is 4 persons per house (Department of census and statics, 2001).

Instead of calculating individual equipment power consumption, it is more practical to use socio-economic factors to predict domestic electricity consumption in many situations. The present study used the number of family members and the income level as inputs for the model building. The maximum prediction accuracy achieved was 90% (for test data) by the "Random Forest" algorithm. But it should be noted that when pre-processing the data, nearly 23% of data was lost. This may lead these models to overfitting, as for further studies, much smaller income ranges are preferred, such as.

$$\text{IRZ} = \frac{MFI}{k} \text{ ; where k is equal or greater than 2} \tag{9}$$

However, these results can further be improved by feeding more data to train the models.

ACKNOWLEDGEMENT

Financial assistance of the World Bank AHEAD-DOR grant is acknowledged.

REFERENCES

Aleksandar, D., Sonja, F., & Ljupco, K. (2016). Deep belief network based electricity load forecasting: An analysis of Macedonian case. *Energy*, 1688-1700.

Amarawickrama, H. A., & Lester, C. H. (2007). *Electricity Demand for Sri Lanka:*. Guildford, UK: Department of Economics, University of Surrey.

Çamurdan, Z., & Ganiz, M. C. (2017). Machine learning based electricity demand forecasting. *2nd International Conference on Computer Science and Engineering, UBMK 2017*, 412-417.

Department of census and statics. (2001). *Census of Population and Housing 2001.* Colombo, Sri Lanka: Department of census and statics.

Department of cenus and statics. (2016). *Household income and expenditure survey 2016.* Colombi, Sri Lanka: Department of cenus and statics.

Ho, T. K. (1995). Random Decision Forests. *3rd International Conference on Document Analysis and Recognition*, (pp. 278-282).

Koehrsen, W. (2018, August 30). *An Implementation and Explanation of the Random Forest in Python*. Retrieved from Towards data science: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76

Minitab. (2013, May 30). *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?* Retrieved from The Minitab blog: https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Nti, I. K., Samuel, A.-a., Michael, A., & Asafo-Adjei, S. (2019). Predicting Monthly Electricity Demand Using Soft-Computing Technique. *International Research Journal of Engineering and Technology (IRJET)*, 1967-1973.

Nti, I. K., Teimeh, M., Nyarko-Boateng, O., & Adekoya, A. F. (2020). Electricity load forecasting: a systematic review. *Journal of Electrical Systems and Information Technology*.

Rajmohan, K., & Weerahewa, J. (2010). Household Energy Consumption Patterns in Sri Lanka. *Sri Lankan Journal of Agricultural Economics*, 55-77.

Saberbari, E., & Saboori, H. (2014). Net-Zero energy building implementation through a grid-connected home energy management system. *19th Conference on Electrical Power Distribution Networks, EPDC 2014* (pp. 35-41). Tehran, Iran: IEEE.

Varoquaux, G. (2013, August 22). *scikit-learn: machine learning in Python*. Retrieved from Scipy Lecture Notes: http://scipy-lectures.org/packages/scikit-learn/index.html