Communications in Applied Sciences
ISSN 2201-7372
Volume 8, **2020**, 1

INFINITY PRESS
www.infinitypress.info

Application of Computational Techniques to Analyse Big Data

Ozoh, P[1]*; Olayiwola, M[2]; Ogundoyin, I[1]

[1]Dept of ICT, Osun State University, Osogbo, Nigeria
[2]Dept of Mathematical Sciences, Osun State University, Osogbo, Nigeria
*Corresponding author: patrick.ozoh@uniosun.edu.ng

Abstract

Computational analysis is a collection of procedures that is used to process large amounts of data with a view of obtaining results based on processed data and as a result, getting their behavioral pattern. The main goal of this research paper is to apply computational techniques to analyse big data, and to analyze the behavioral pattern of such data. To effectively evaluate estimated data, a simulation is carried out to accurately model actual data and at what period of time. The simulation, based on real data, is developed to estimate actual data. A goodness-tests which consist of Chi-Square (χ^2) goodness-test is computed to ensure that data estimates are reliable and adequate.

Keywords: Computational techniques, behavioral pattern, big data, simulation, evaluation, goodness tests.

1. INTRODUCTION

Big data is data sets that are so big and complex that traditional data processing application software are inadequate to deal with them. Big data refers to the use of predictive analytics, user behavior analytics that extract value from data. Analyses of data sets are able to find new correlations to identify business trends, applied to medical healthcare, improve security, financial institutions, and so on. Big data is highly used by scientists, business executives, medical practitioners, governments, etc., who regularly face difficulties with large data-sets in informatics. Some of these difficulties can be found in areas in meteorology, genomics, complex simulations, biology, and environmental research.

This study identifies and applies accurate computational techniques to increase reliability in estimating big data. Modern day planning is based on deriving precise estimates for developing models. The analysis and modeling of real world problems are important features for performing decision making in diverse applications. This ranges from manufacturing applications, telecommunications, construction, military applications, health applications, logistic, transportation distribution, to mention just a few. In order to estimate predictive values, it is expedient to apply an appropriate machine learning technique that will produce accurate estimates.

Computational techniques are used for the discovery of patterns and relationships in sets of data. The fundamental goal of any computational techniques is to discover meaningful or non-trivial relationships in a set of data and produce a generalization of these relationships that can be used to interpret new, unseen data.

As a result, the contributions to research of this paper are as follows:

1. Capturing of data used for analyzing big data to get insights and trends

2. Computing estimates and forecasts for real world data

3. Evaluation of models obtained from estimated data

2. LITERATURE REVIEW

Developments in the field of computational analysis is often parallel or follow advancement in those fields whereby statistical computing is applied. It is because computational analysis often address particular applied decision problems, methods developments is consequently motivated by the search to better decision making under uncertainties. In order to perform reliable decision making, it is expedient to accurately model real life problems in diverse applications in order to accurately estimate actual data, otherwise inappropriate models and poor estimates may occur (Ozoh et al., 2018). The paper presented that models are essential in providing support for businesses processes, systems and dealing with complex problems. The development of appropriate models for planning and management is a tool for improving efficiency in real world problems. The integrated grey model with multiple regression model (IGMMRM) was applied to modeling of data, in comparison with Grey model (GM) and multiple regression method (Wang & Xia, 2009). The modeling techniques were assessed using *relative error* (RE), *mean absolute error* (*MAE*), *root mean square error* (*RMSE*), and *mean absolute percentage error* (*MAPE*) to evaluate the accuracy of the models. The study suggests that the performance of IGMMRM was higher than the other two models based on historical data.

A paper by Widén and Wäckelgård (2010) presented a modelling framework for the generation of high-resolution series of data. The model generated synthetic estimates. The activity-generating model based on non-homogeneous Markov chains were converted to an extensive empirical time-use data set creating a realistic spread of activities over time, down to a 1-min resolution. Artificial neural network (ANN) was

described in Damak (2011) as a hidden-layer feed-forward network technique and it's a widely used technique for time-series modelling and forecasting. The paper described that the technique is based on pattern recognition, and able to forecast for non-linear models. Neural networks are similar to the least square estimation technique and can be viewed as an alternative statistical approach to solving least squares problems (Chen et al. 2014). The paper presented an artificial neural network-based short-term load forecasting technique for estimating data. The ANN technique utilized a combination of the three layer feed-forward neural network and a back-propagation training technique. An artificial neural network transport energy demand model was developed by Murat (2006) using gross national product (GNP), population, available historical energy data and transport related indicators. The model was obtained using a feed-forward neural network, trained with the back propagation algorithm.

A novel regression technique, evolution local kernel regression (ELKR) was introduced by Akole and Bongulwar (2011). The research paper applied artificial neural network, which utilized historical data. This technique is based on local Nadaraya-Watson estimates with independent bandwidths distributed in data space. The model utilized the covariance matrix adaptation evolution strategy (CMA-ES), a stochastic method for solving non-linear numerical optimization problems. The performance evaluation statistics of the developed prediction model were computed using *mean absolute percentage error* (*MAPE*), *mean square error* (*MSE*), *root mean square error* (*RMSE*) and *percentage error* (*PE*). The result of the research indicated that values of the evaluation statistics for ANN technique were low compared with multiple regression technique, which shows ANN to be more accurate and effective than multiple regression technique for load and price forecasting. Goh (1998) employed the univariate Box-Jenkins approach, multiple log-linear regression and ANN techniques to compare forecasting accuracy of residential consumption demand. The forecasting accuracy of the methods was

compared using percentage errors for the three techniques. The study indicated the superiority of ANN to other techniques, since it has the lowest *mean absolute percentage error* (MAPE) value.

Zhang et al. (2010) applied ANN technique to estimate data. Using two deterministic chaotic time series generated by the logistic map and the Glass-Mackey equation, the paper designed the feed-forward neural networks to predict such dynamic nonlinear systems. Their results show that ANN can be used for modelling and forecasting nonlinear time series with very high accuracy. ANN, which embodies a large degree of uncertainty useful for predicting historical data, offers a great deal of promises. The analysis of a prediction model built on ANN based on learning, flexibility and real-time response was illustrated by Yedra et al. (2014). Previous research on estimating models also utilized the ANN (Marvuglia, 2012)

3. EXPERIMENTAL DESIGN

This section discusses data collection methods, findings from occupants' attitudes towards energy conservation which is conducted by distributing questionnaires to electricity consumers. This section also discusses the simulation used in this study which involves analyzing the data collected using Chi-Square (χ^2)data analysis goodness-test procedure. .

In this study, data is collected from a selected building. Population and sample must be clearly identified for which inference would be made where all requirements of sampling and experimental design must be satisfied. The research samples are given as follows:

Population:

- All buildings consisting of living rooms, bedrooms and toilets

Sample: The selected building consists of

➢

- Living room
- Bedrooms
- Toilet

The research sample is given in Figure 1.

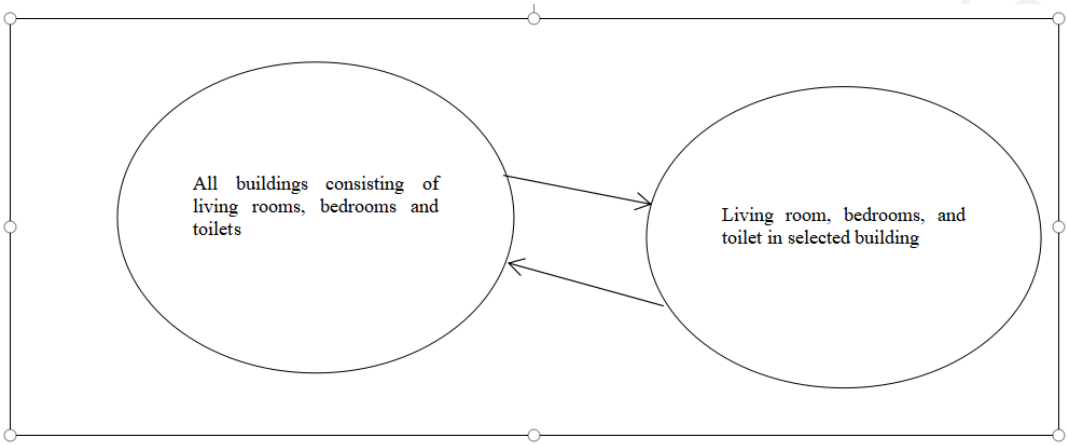

Figure 1: Research Sample for Data Collection

The samples are picked using random sampling method where all subjects that is in the population have equal chances of been chosen. For this research, the living room, bedrooms, and toilet are numbered in the selected building. Then, Excel is used to generate random numbers to create a group of sample. Data collection was carried out for 21 days, randomly, between 20 Feb 2018 – 22 Jun 2018 on weekdays except during public holidays. Observation technique was used to calculate the duration of time electricity is unnecessarily used. For example, in the toilet, when person 1 enters the toilet for about 5 minutes, electricity is not switched off after he leaves the toilet. After another 5 minutes, person 2 walks in, the duration for the electricity consumed is calculated from the time person1 leaves the toilet and the time person 2 enters. The design for data collection and electricity wastage is shown in Table 1.

Table 1. Observation Data

	Arrival time	Departure time	Duration (mm:ss)	Switch off/ left on	Is there another person inside? (Y/N)	Wastage (hh:mm:ss)
Person 1	8:35:01 AM	8:38:56 AM	3 :55	On	N	00:04:27
Person 2	8:43:29 AM	8:48:00 AM	5 :29	On	Y	00:00:00

Person 3 occupies the living room from 9a.m to 11p.m. Later, the TV and lights are not switched off, also the PC is not changed to sleep mode. The next day, person 4 occupies the living room at 2p.m. The duration for the room to be vacant will be taken into account. The room is checked every 2 hours to see whether the electrical equipments are still on the same condition as they are left. Table 2 shows the design of the data collection for observation data.

Table 2. Observation Data for Living Room

Time	Monday
8-9	
9-10	
10-11	
11-12	On
12-1	On
1-2	On
2-3	On
3-4	On
4-5	On

A survey is conducted by distributing electricity consumers attitudes towards energy conservation (ATEC) questionnaires to study occupants' behavior. The occupants' attitude towards energy conservation used in this study contains 15 questions with multiple choices and in likert scale format. Likert scale is anchored at both ends (1= Strongly Disagree and 5= Strongly Agree). ATEC is composed of two parts; (i) 7 questions on top was to measure the respondent behavior pattern of the energy usage, and (ii) 8 questions with likert scale to measure the respondent awareness of the energy conservation. The questionnaires were distributed to two selected samples of the Oke Baale community of the Osun State University, comprising staffs and students of the university. The electricity consumers used for the survey are selected randomly based on their availability.

4. RESULTS AND DISCUSSION

The occupants' attitudes towards energy conservation questionnaires were analyzed separately between staffs (40 questionnaires) and students (60 questionnaires). The findings reflect Osun State University staffs and students behavior pattern towards energy conservation. Excel spreadsheet was used to plot the graphs. The results obtained from analyzing then questionnaires are given as follows:

Figure 2 shows how staff and students leave their personal computer (PC) or laptop when not in used. Most of the staff shut their PC or laptop when they are not using it about 21 people or 52.5% of them. 14 people or 35% of them put the PC or laptop in standby mode and only 5 people or 12.5% left the PC or laptop on even when unused. There are 25 people or 41.7% of the students who put their PC or laptop in standby mode when they not using it. Meanwhile, there are 23 people or 38.3% of them who shut down their PC or laptop and there only 12 person or 20% of them who left their PC or laptop on.

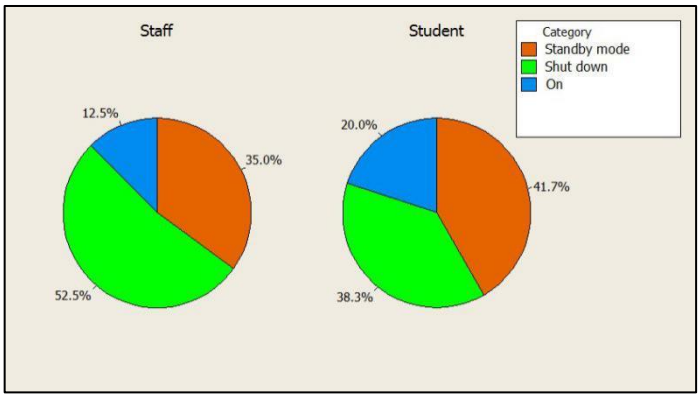

Figure 2: PC or Laptop Mode when not in use

Figure 3 shows the percentage on the average frequency of personal computer or laptop left on when staffs and students leaving their room. The most frequent of average time that staffs left their PC or laptop on is more than an hour which is 14 people or 35% of them. Then there are 15% of staffs or 6 people chose 10 to 15 minutes and 5 to 10 minutes. Lastly, only 10% of staffs or 4 people stated they never left their PC or laptop on. 41.7% of students or 25 people stated the average frequency they left their PC or laptop on while leaving the room is more than an hour. 28.3% of students or about 17 people chose 15 to 30 minutes.10% of students or 6 people chose 10 to 15 minutes and 5 to 10 minutes as their average time they left their PC or laptop on. 6 persons or 10% of students stated they never left their PC or laptop on while leaving the room.

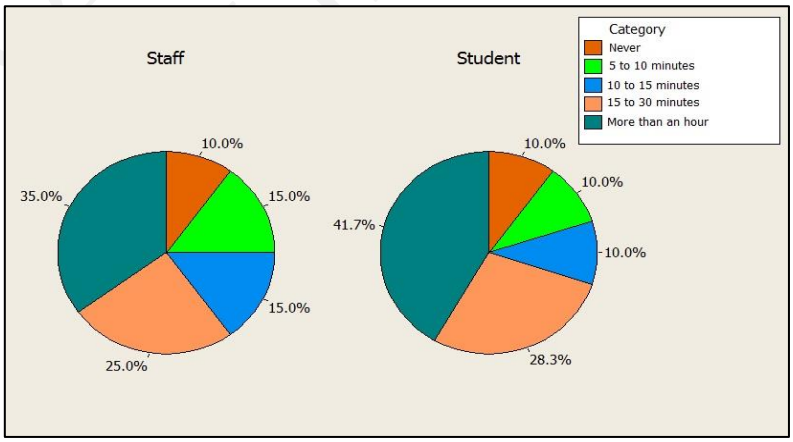

Figure 3: Frequency of PC or Laptop Left on when leaving Room

Figure 4 shows the frequency lights are switched off when the last person leaves the room. For staffs, there are 62.5% of them or 25 people that switch off the lights when they are the last person to leave the room. There are 7.5% or 3 people that rarely switch off the lights when they are the last person to leave the room. 1 person or 2.5% never switch off the lights. The frequency of students for every time is 36.7% or 22 people. The frequency for rarely switch off lights is 7.5% or 3 people, while 1 person or 2.5% never switches off lights.

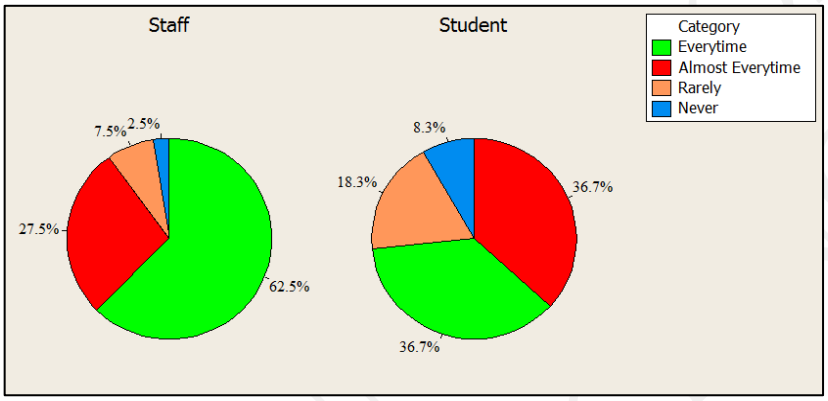

Figure 4: Frequency of Switching off Lights when the last Person leaves Room

4.1. Simulation Results

The Chi-Square (χ^2) data analysis goodness-test procedure is applied in this study and is given in Table 3. Chi-Square (χ^2) is given as:

$$\chi^2 = \frac{(o_i - np_i)^2}{np_i}$$

Where,

O_i = observed values, n = sample size, k is the number of class intervals, and

$$p_i = \frac{1}{k}$$

Table 3. Chi-Square (χ^2) Data Analysis Procedure

Research Hypotheses	Data Source
Ho: the random variable is exponentially distributed *H1*: the random variable is not exponentially distributed	Observation data for toilet at ground floor Monday- Friday (8 AM – 1PM)

The computed results of the Chi-Square (χ^2) test for Monday, Tuesday, Wednesday, Thursday and Friday for toilet together with their respective table $\chi^2_{0.01}$ at α = 0.01 are given in Table 4.

Table 4. Chi-Square Goodness-of-Fit Test for Inter-Arrival Time for Toilet on Monday to Friday, February 25(in minutes)

Days	Computed Chi-Square (χ^2)	$\chi^2_{0.01}$
Monday	1.238	15.5
Tuesday	13.024	15.5
Wednesday	6.217	15.5
Thursday	14.558	15.5
Friday	14.229	17.5

On Monday, the degrees of freedom is given by *k-s*-1 = 10-1-1 = 8. At α = 0.05, the table value of $\chi^2_{0.05,8}$ is 15.5. Since $\chi^2_0 < \chi^2_{0.05,8}$, thus we fail to reject the null hypothesis. Note that the value of $\chi^2_{0.01,8}$ is 20.1, so the null hypothesis would also not be rejected at level of significance α = 0.01. Thus this test gives us no reason that our data are poorly fitted by the exponential distribution.

For Tuesday, the degrees of freedom is given by *k-s*-1 = 10-1-1 = 8. At α = 0.05, the table value of $\chi^2_{0.05,8}$ is 15.5. Since $\chi^2_0 < \chi^2_{0.05,8}$, thus we do not reject the null hypothesis. Note that the value of $\chi^2_{0.01,8}$ is 20.1, so the null hypothesis would also not be rejected at level

of significance α = 0.01. Thus this test gives us no reason that our data are poorly fitted by the exponential distribution.

For Wednesday, the degrees of freedom is given by k-s-1 = 10-1-1 = 8. At α = 0.05, the table value of $\chi^2_{0.05,8}$ is 15.5. Since $\chi^2_0 < \chi^2_{0.05,8}$, thus we failed to reject null hypothesis. Note that the value of $\chi^2_{0.01,8}$ is 20.1, so the null hypothesis would also not be rejected at level of significance α = 0.01. Thus this test gives us no reason that our data are poorly fitted by the exponential distribution.

On Thursday, the degrees of freedom is given by k-s-1 = 10-1-1 = 8. At α = 0.05, the table value of $\chi^2_{0.05,8}$ is 15.5. Since $\chi^2_0 < \chi^2_{0.05,8}$, thus we failed to reject the null hypothesis. Note that the value of $\chi^2_{0.01,8}$ is 20.1, so the null hypothesis would also not be rejected at level of significance α = 0.01. Thus this test gives us no reason that our data are poorly fitted by the exponential distribution.

For Friday, the degrees of freedom is given by k-s-1 = 10-1-1 = 8. At α = 0.025, the table value of $\chi^2_{0.025,8}$ is 17.5. Since $\chi^2_0 < \chi^2_{0.025,8}$, thus we failed to reject null hypothesis. Note that the value of $\chi^2_{0.01,8}$ is 20.1, so the null hypothesis would also not be rejected at level of significance α = 0.01. Thus this test gives us no reason that our data are poorly fitted by the exponential distribution.

4.2. Model Building

A few assumptions are applied which are as follows:

(i) Only one person enters the toilet at any point in time.

(ii) The distribution of arrivals and duration times is the same for every week.

(iii) We also assume that there is no wastage after Saturday 8AM as the sample size for the data is small.

For the simulation of the toilet, heuristic model is presented and evaluated that effectively satisfies the problem. Heuristic model is a situation in which all relevant alternatives, their consequences, and probabilities are known, and where the future is certain, so that the optimal solution to a problem can be determined (Gaissmaier, 2011). Thus, the following variables are investigated:

λ = the mean of user arrival rate

μ = duration rate

The arrival rate (λ)and the duration rate (μ) are used in the simulation to generate the sequence of the arrival times. The results are given in Table 5.

Table 5. Summary of Arrival Rate and Duration Rate for Toilet

	Time interval	λ	μ
Monday	8 AM – 1PM	0.097	3.282
Tuesday	8 AM – 1PM	0.096	3.908
Wednesday	8 AM – 1PM	0.08	3.526
Thursday	8 AM – 1PM	0.089	3.910
Friday	8 AM – 1PM	0.096	3.682

4.3. Verification and Validation of Simulation Models

The validation test is conducted using historical input data to ensure that model will duplicate closely as possible the important events that occurred in the real system (Banks et al., 2000). The data that will be used is the duration of the real data (Monday 8AM-11AM) and the model output given in Table 6.

A paired-test was conducted to test H_0: $\mu_d = 0$, or equivalently, H_0: $E(Z_1) = E(W_1)$, where Z_1 is the duration from the real system and W_1 is the duration predicted by simulated model. Let the level of significance be α= 0.05. Using the results in Table 6, the test statistic is

$$t_0 = \frac{\bar{d}}{s_{d/\sqrt{K}}} = 0.723/(\frac{10.428}{\sqrt{87}}) = 0.647$$

The critical value is $t_{\alpha/2,K-1} = t_{0.025,86} =$ 2.00. Since $|t_0| = 0.0647 < t_{0.025,67}$, the null hypotheses cannot be rejected on the basis of this test; that there is no inconsistency is detected between system output and model prediction. As a result, the simulated model is proven to be valid and we assume that the model is adequate to simulate the real world situation.

Table 6. Validation of the Simulation

Input Data Set j	Real-System output $Z_{i,j}$	Model Output $W_{i,j}$	Observed Difference d_j	Squared Deviation from Mean $(d_j - \bar{d})^2$
1	2.383	0.813	1.57	0.717
2	0.967	5.565	-4.598	28.318
3	1.9	1.057	0.843	0.014
4	1.833	1.753	0.08	0.414
5	2.333	0.621	1.712	0.977
6	0.65	0.834	-0.184	0.824
7	5.85	0.252	5.598	23.761
8	2.8	0.287	2.513	3.202
9	2.617	2.727	-0.11	0.695
10	3.117	0.892	2.225	2.255
11	2.4	4.031	-1.631	5.544
12	5.817	8.678	-2.861	12.848
13	5.5	0.735	4.765	16.334
14	3.05	1.724	1.326	0.363
15	1.467	1.826	-0.359	1.172
16	1.217	0.249	0.968	0.06
17	2.333	1.23	1.103	0.144
18	2.65	1.199	1.451	0.529
19	6.1	0.377	5.723	24.995
20	1.633	0.538	1.095	0.138
21	3.533	1.369	2.164	2.075
22	6.7	4.226	2.474	3.064
23	2.117	3.551	-1.434	4.655
24	1.117	0.3	0.817	0.009
25	4.45	8.835	-4.385	26.096
26	6.5	1.775	4.725	16.012
27	0.75	4.438	-3.688	19.461
28	4.417	0.054	4.363	13.246
29	2.15	0.843	1.307	0.341
30	4.433	4.292	0.141	0.339
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
87	1.2	0.665	0.535	0.036
			$\bar{d}$= 0.723	S_d^2= 10.428

4.4. Analysis of Electricity Wastage

Based on collected data, further analysis is needed for the calculation of estimating the cost of energy wasted in the selected building and is given as follows:

(i) Electricity wastage consisting of the sum of the total amount of hours spent on each electricity appliances, and finding the average of the sum for each appliance per day.

(ii) The electricity appliances used in the selected household are identified. Then, the electricity consumption (wattage) for the electricity appliances are searched online, and used to estimate the electricity wastage. In average, electricity consumption of the electricity appliances on wastage per day is calculated. The electric wastage is calculated using Table 1.

(iii) The cost of energy wasted is calculated using electricity tariff of the Transmission Company of Nigeria (TCN), which is N31.26 per kwh.

From the simulation process given in Table 1, the estimate of electricity wastage for the selected household is obtained and is displayed in Table 7.

Table 7. Summary of Estimated Electricity Wastage based on Simulation

Day	Total wastage /hour	Total watt/hour	Total kw/hour
Monday	304.663	76775.076	76.775
Tuesday	334.450	84281.400	84.281
Wednesday	383.306	96593.112	96.593
Thursday	339.707	85606.164	85.606
Friday	301.375	75946.500	75.947
		∑	419.202

From Table 7, the cost of electricity wastage for the household is 419.202× N31.26 = N 13, 104. 25 (About US$ 36.40).

5. CONCLUSION AND FUTURE WORK

In summary, the objectives of this research are achieved. The first objective of this research is to investigate the occurrence of electrical wastage in the selected household. This was achieved by collecting data in the building. In achieving the second objective of identifying energy consumers attitude towards energy conservation, questionnaires were distributed to them. Lastly, the third objective is achieved by doing a simulation to calculate the estimated cost of the electricity wastage in the building.

In this study, the estimated amount of electricity wastage in the building is obtained. Their is consumer awareness towards energy saving, and also the importance of the energy saving, but consumer behavior towards energy saving needs to be improved. Hence the need to consider implementing an appliance that can help in reducing electricity wastage, such as a light sensor. By implementing a sensor, the cost of electricity wastage can be reduced.

Based on this research, the future works to this research can be expanded as follows:

(i) To make more solid simulations that can represent the real-world situation perfectly

(ii) Identify the electricity wastage in additional buildings

References

[1] P. Ozoh, S. Abd-Rahman, M. Olayiwola. (2018). Developing Predictive Models using Typical Machine Learning and Computational Techniques, Analele Universităţii "Tibiscus", Timişoara, vol. 16, no. 2, pp. 82-85.

[2] F. Wang, X. Xia. (2009). Integration of Grey Model and Multiple Regression Model To Predict Energy Consumption. IEEE Proceedings, pp. 194–197.

[3] J. Widén, E. Wäckelgård. (2010). A High-Resolution Stochastic Model of Domestic Activity Patterns and Electricity Demand. Applied Energy, 8vol. 7, no. 6, pp. 1880–1892.

[4] S. K. Damak. (2011). Applications of Two Identification Methods For an Electric Distribution System, Journal of Automation & Systems Engineering, vol. 4, pp. 176–184.

[5] L. Chen, X. Xu, L. Yao, Q. Xu. (2014). Study of a Distribution Line Overload Control Strategy Considering the Demand Response. Electric Power Components and Systems, vol. 42, no. 9, pp. 970–983.

[6] Y. S. Murat, H. Ceylan. (2006). Use of Artificial Neural Networks for Transport Energy Demand Modelling. Energy Policy, vol. 34, pp. 3165–3172.

[7] M. Akole, B.T. Bongulwar. (2011). Predictive Model of Load and Price for Restructured Power System using Neural Network, International Conference on Energy, Automation, and Signal (ICEAS), pp. 1–6.

[8] F. Goh. (1998). Forecasting Residential Construction Demand in Singapore: A Comparative Study of the Accuracy of Time Series, Regression and Artificial Neural Network Techniques. Engineering, Construction and Architectural Management, vol. 5, no. 3, pp. 261–275.

[9] G. P.Zhang, P. Areekul, S. Member, T. Senjyu, H. Toyama. (2010). A Hybrid ARIMA and Neural Network Model for Short-Term Price Forecasting in Deregulated Market. IEEE Transactions on Power Systems, vol. 25. no. 1, pp. 524–530.

[10] F. Yedra, M. Diaz, A. Nieto. (2014). A Neural Network Model for Energy Consumption Prediction of CIESOL Bioclamitic Building, Advances in Intelligent Systems and Computing, vol. 239, pp. 51-60.

[11] A. M. Marvuglia. (2012). Forecast, Using Recurrent Artificial Neural Networks to Consumption, Household Electricity, Energy Procedia, vol. 14, no. 1.