(RESEARCH ARTICLE)

Check for updates

Application of agglomerative hierarchical clustering and logistic model development for assessing solar energy acceptability as an alternate energy option

Koyejo Oduola * and Zorbarile Atukomi

Chemical Engineering Department, University of Port Harcourt, Nigeria.

Abstract

This paper is focused on the assessment of acceptability of solar energy as an alternate efficient energy management option using Agglomerative Hierarchy Cluster (AHC) and logistic regression modelling approach. The study population includes randomly selected shop-owners and residential occupants within the Port Harcourt city in Rivers State, Nigeria. The collected data sets were subjected to AHC analysis using a statistical package XLSTAT 2016 version 4.6. The central object identified from the application of AHC with respect to the sampled shop-owners and residential occupants as pertaining to the acceptability of solar energy as an alternate efficient energy management option was centered around the financial implication of energy generation and the political influence of the government solar energy policies for energy generation. Finally, logistic regression modelling approach was applied into developing a predictive model for the probability of general acceptance (variable 'yes') of solar energy as an effective energy management system. From the developed model the chance of acceptance of a solar energy management system is 1% with 59.5% rejection from the study population while it is 99% with an unawareness level of 40.51% from the study population.

Keywords: Acceptability; Agglomerative hierarchy clustering; Logistic regression modeling: Solar energy

1. Introduction

Given the huge potential for solar power generation (and other forms of renewables), the whole world including Nigeria has been making considerable efforts to diversify its energy mix (with less emphasis on thermal generation) and invest heavily on research and development on the exploitation of solar and other renewable energy technologies. In terms of a legal, regulatory and tariff framework, a lot has been done in terms of policy and regulation [1, 2, 3, 4, 5, 6].

Procurement, establishment/ installation as well as advancement of a solar system in a megacity could be hindered with issues such as economic viability, inadequate research and innovation into solar energy development, non-compliance to existing buildings and other power distribution challenges and also the unawareness/unacceptability of the general public who are not familiar with the increasing contribution of solar based energy to the total electricity source. Majority are unaware of the fact that solar PV systems are in measured frame and can be designed and arranged in series and in parallel to accomplish the coveted power yield. Solar thermal can deliver heat by consolidating temperature and mass (water or hydrogen et cetera) running into Kilowatt or Megawatt to run a turbine that can create an equivalent measure of electric power as the current ordinary power supply [2, 4].

The problem of storing large amount of solar energy after it has been converted to electrical energy is a huge challenge yet to be overcome before solar energy becomes a major contributor to the world energy grid [6]. A major infrastructure investment will be necessary for such a storage system to be possible. Also, transporting energy from

* Corresponding author: Koyejo Oduola
Chemical Engineering Department, University of Port Harcourt, Nigeria.

where it is produced to where it is needed is another huge challenge to be overcome. A new high- voltage, direct-current (HVDC) power transmission backbone would have to be built using Direct Current for this to be possible [3, 5].

Another serious setback to the solar energy program is ignorance of the benefits of this technology. Awareness of the opportunities offered by solar energy and its technology is low among members of the public and private sector stakeholders. This lack of information and awareness creates a market distortion that results in higher risk perception for potential renewable energy projects.

Solar irradiation which varies throughout the entire day and affects the efficiency and output of PV cells is another issue being considered in using solar cells. Increase in solar irradiance increases the PV module efficiency because the high number of photons hitting the module increases and many electron-hole pairs are formed which will produce more current. During the night, solar irradiation is zero hence PV cells will have zero output at night. A simple way to solve this problem is to incorporate another renewable energy source such as wind energy with the solar PV modules so that they will deliver the required power at night. Electricity storage in batteries is also useful to make electrical energy available during these periods when solar irradiation is low or not available [7, 8, 9].

Several attempts have therefore been made to develop models associated with solar energy. There are various illustrations where GIS has been utilized to help the arranging and planning procedure of sustainable power source. Particularly, the distinguishing proof of appropriate spots for wind and solar based homesteads, pump stockpiling hydroelectricity, as well as the mapping of locating of sustainable power source assets [10]. These examinations utilize geospatial information and data ashore, utilize, rise, structures and foundation. A large portion of the investigations tend to address the potential energy supply without considering the level of demand. Nevertheless, it is very informative, such an examination requires more information and more propelled investigative strategies. For example, a study carried out by Kucuksari, proposed a system that consolidates GIS, scientific optimization and stimulation with a specific end goal to locate the ideal size and the ideal area of photovoltaic plants for the campus environments [10].

The GIS module serves for distinguishing proper housetops and their photovoltaic panel capacity. In any case, this approach is simply in view of static geodata (for this situation, light detective and running (LiDAR) information) and does not represent dynamic land factors, for example, climate conditions, or solar based radiation and wind [11].

However, this approach could be the reason for a more complete structure that likewise incorporates extra spatially changing sustainable sources, for example, geothermal energy or wind sources. In locales where the potential for growing district heating (DH) systems veers in various territories, the monetary expenses of heat production, transmission and distribution are hard to assess. Along these lines, considering the geographic segment utilizing geospatial techniques and GIS is a basic part in finding the limits to which such a development is financially plausible. Despite the fact that it was not specified by the authors, the most proficient utilization of energy stockpiling innovation, for this situation, heat energy stockpiles, likewise relies upon spatial parameters, for example, the separation to the following piece or to singular houses. The ideal utilization of energy stockpiles in the production of heat, the transmission of heat and the distribution of heat may fundamentally have impact on the general expenses on the areas of power supply [11]. They concluded that the heat supply choices ought to be founded on the spatial position of the demand of heat and the attributes of the nearby DH area, which verifiably supports the basic need of GIS [12].

Solar radiation data are required for a number of solar thermal and Solar photovoltaic applications like solar power generation, solar heating, cooking, drying and solar passive design of building. The measured solar radiation data are not available for most of the sites due to high cost, and maintenance of the measuring instruments. As such, various empirical models have been used to predict monthly mean daily solar radiation all over the world [12]. The Artificial neural networks (ANNs) are used to solve a number of scientific problems. It has the capability to approximate any continuous non-linear function to arbitrary accuracy. A multi-layer feed-forward neural network can approximate a continuous function due to its robustness, parallel architecture and fault tolerance capability. In past years, ANN models are used by a number of researchers to estimate solar radiation and concluded that ANN model are proven to be superior to other empirical regression models [13].

The ASHRAE clear sky model is commonly used as a basic tool for solar heat load calculation of air conditioning systems and building designs among the engineering and the architectural communities in Thailand. Recently, Joeijoo and Sorapipatana assessed the accuracy of the ASHRAE model in the northern part of Thailand [14]. They found that the ASHRAE model is considerable over estimation for direct radiation, and under estimation for diffuse radiation. The reason for large errors in the prediction of the ASHRAE model stemmed from the assumptions of the standard atmospheric condition used in the model itself, which hypothesized the clear sky as for a typical nonindustrial mid

latitude atmospheric condition in USA. As a result, it gives large deviations for predictions in the tropical climate as in Thailand.

This study has been centered on assessing the general awareness and acceptability of solar energy as an alternate energy management system among residence. The study area with respect to this study was Port Harcourt city with the study population consisting of two sampled groups of respondents (shop-owners and residential occupants). The major instrument used for data collection was structured questionnaires. The collected data includes both primary and secondary data. We have recently reported a part of the research which employed for the analysis Kolmogorov-Smirnov distribution test and one-way single factor analysis of variance [15]. Herein, Agglomerative Hierarchy Cluster (AHC) and logistic regression model at 95% confidence level were the statistical analyses applied on the collected data sets. The criterion for model acceptance was based on the resultant likelihood function. An attempt has been made to develop a model for predicting the probability of acceptance of solar energy as an alternate source within Port Harcourt city applying the principle of logistic regression analysis.

2. Methodology

2.1. Experimental Design/Data Collection

This study has employed an opinion survey design, as a great way for collecting information and explore relationship between different variables when human beings are the units of analysis. The research process involves data gathering, tabulating, description, analysis and interpretation. The required data were obtained by direct observations, distribution of questionnaires to sampled respondents and inventory format.

The study area of this research is the largest and capital city of Rivers State, Nigeria, Port Harcourt. Rivers state has a population of over 5 million, and its capital city is of vast economic significance as the center of Nigeria's oil and gas industry. The population of this study was limited to shops owner/market sellers and general residential occupants within Port Harcourt city. The general demographic details with respect to the sampled respondents include their level of education; sex, age group, and duration of occupancy. Details on the sampling and sampling technique used as well as the method of data collection have been reported in a recent work [15]. Typical questionnaire employed for the research is shown here also in Table 1.

2.2. Method of Data Analysis

The statistical package employed for data analysis was XLSTAT 2016 version 4.6. Agglomerative Hierarchical Clustering (AHC) and Logistic Regression Analysis have been applied for analyses of the collected data.

Agglomerative hierarchical clustering (AHC) is an important and inherent process in unsupervised machine learning [16]. It is a statistical process that pairs and merges a group of data clusters from bottom up during an upward movement of observation data clusters. In agglomerative hierarchical clustering (AHC), observations which are in clusters in their original state starts to pair as the, progress upwards. These paired clusters are then merged in an upward hierarchical structure in sub clusters and sub-sub clusters formation until a single cluster containing all required documents is formed. The benefits of agglomerative hierarchical clustering (AHC), is that informative data display can be revealed from ordering of the single data objects in the clusters. Also, discoveries are made possible from the handling of smaller data clusters that are generated during the course of the hierarchical agglomeration.

The typical questions employed for data collection have been listed below.

Code   Questionnaire Parameters

QP1   Are you aware of solar energy?
QP2   When compared to other sources of energy will solar energy be more environmentally friendly?
QP3   Would you love to install your own solar energy equipment?
QP4   At home or in your office/shop what alternative energy source do you apply?
QP5   What is your weekly budget like for the maintenance/service of the fuel generator?
QP6   In your opinion, could solar serve as a great substitute for the fuel generator?
QP7   Do you accept that solar energy is the most promising renewable energy source owing to its comparative limitless potential?
QP8   Could solar energy be more cost effective in the long run, despite its capital intensiveness?

QP9   Would you recommend that government put policies in place towards subsidizing the installation cost of solar energy source rather than petrol and kerosene?
QP10   Will solar energy be more ideal option as a result of its abundance and accessibility?
QP11   Do you think that solar energy is the same source of energy as nuclear energy?

Agglomerative Hierarchy Cluster (AHC) begins with each variable signifying an individual cluster. These are then subsequently merged according to their similarity. Initially, the two most similar clusters (usually those with the smallest distance between them) are merged to form a new cluster at the bottom of the hierarchy. In the next step, another pair of clusters is merged and linked to a higher level of the hierarchy, and so on [16]. The similarity proximity type employed in this study was with respect to Kendal correlation coefficient and the agglomerative method used was the unweight pair-group average approach.

AHC as an iterative classification method follows simple procedures.

- Estimation of the dissimilarity between the N objects.
- Identified two objects which when clustered together minimize a given agglomeration criterion will be clustered together to create a class made up of these two objects.
- Evaluation of the dissimilarity between this resulting class and the N-2 other objects using the agglomeration criterion.
- The two objects or classes of objects whose clustering together minimizes the agglomeration criterion are then clustered together.

This process continues until all the objects have been clustered. These successive clustering operations produce a binary clustering tree (dendrogram), whose root is the class that contains all the observations. This dendrogram represents a hierarchy of partitions. It is then possible to choose a partition by truncating the tree at a given level, the level depending upon either user-defined constraints (the user knows how many classes are to be obtained) or more objective criteria.

Logistic Regression is a useful mathematical modeling approach for analyzing the relationship between data that includes categorical response variable such as the presence or absence of a disease, the acceptance or rejection of a system or opinion to a dichotomous dependent variable [17].

Kleinbaum and Klein's [18] Equation (1) presents the logistic function which describes the governing equation of logistic regression model.

$$f(z) = \frac{1}{1 + e^{-z}} \quad (1)$$

This function ranges between 0 and 1, thus, the model is designed to describe a probability which is always some number between 0 and 1. Equation 1 can further be presented by (2).

$$P(y \mid x_1) = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}} \quad (2)$$

Equation 2 can further be expressed in terms of its logit value as (3).

$$\log it(p) = \ln\left\{\frac{\Pr(y=1)}{\Pr(y=0)}\right\} = \beta_0 + \beta_1 x_1 \quad (3)$$

Where p with respect to this study is the probability of acceptance of solar energy as an alternate source of energy. With respect to the collected data, the 'yes' response was taken as the dependent variable while the 'no' response was taken as the independent variables. The 'undecided' response was not used in the data modelling as they are neither here nor there. The collected data set with respect to 'yes' and 'no' with the exception of undecided' data which was not used in the model development were converted to binary data with 'yes' to 1(probability of acceptance) while 'no' equal to zero (probability of unacceptance)[19].

To ensure the validity of the research instrument for this study, a self-structured questionnaire was submitted to the researcher's supervisor and three other experts in the Faculty of Engineering, University of Port-Harcourt, their comments, suggestions and modification of the instrument were used in coming up with the final draft. This was to ensure content validity of the questionnaire.

The reliability of the questionnaire for this study was obtained through the test-retest method. Test-retest reliability is the degree to which scores are consistent over time. It indicates score variation that occurs from testing session to testing session as a result of errors of measurement. Twenty (20) respondents outside the sample of study were administered with the instrument. These respondents consist of twelve (12) shop owners and eight (8) residential occupants. After an interval of three weeks, fresh copies of the questionnaire were re-administered on the same respondents. The reliability co-efficient was established at 0.60 using Pearson correlation test.

3. Results and discussion

The data collected as a result of this study were both primary and secondary data. Figures 1 and 2 present the dendrogram from the application of Agglomerative Hierarchy Clustering on the collected data set with respect to the sampled shop-owners and residential occupants, respectively. Also, Tables 1 and 2 represent the resultant clusters of the collected data sets with their central object for the sampled data sets (Shop-owners and Residential Occupants). Finally, Figures 3 and 4 present the profile plot of the clustered responses with respect to the assessment of the study population on the acceptability and awareness of solar energy as an alternate energy management system and source.
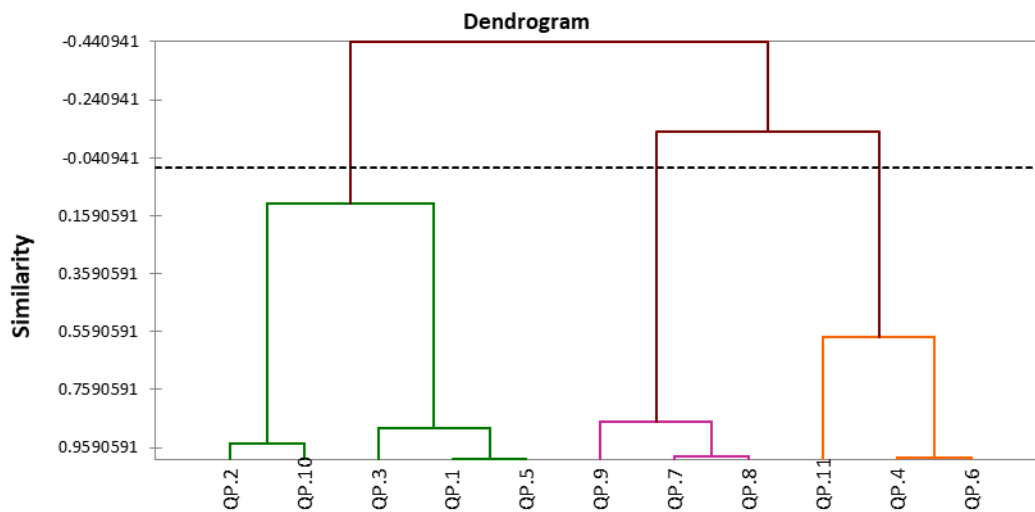

Figure 1 Dendrogram with respect to AHC output on Shop-Owner data set

Analysis of the data collected on the responses of the two sets of data both from Shop-owners and Residential Occupants by applying Agglomerative Hierarchy Clustering (AHC) was performed with the objective to identifying the central opinions or idea of the various respondents, hence identify the differences between their responses and acceptability of solar energy as an alternate source of energy for effective management option. As seen the resultant dendrogram from the application of AHC in Figure 1, the responses of the shop-owners as pertaining to the subject matter were clustered into three major classes with questionnaire parameters QP2, QP10, QP3, QP1 and QP5 belonging to Class 1, while questionnaire parameters QP11, QP4 and QP6 belong to Class 2 and finally questionnaire parameters QP9, QP7 and QP8 belong to Class 3. It is interesting to note that from Table 1, the central object in class 1 was QP5 while class 2 was QP6 and class 3 was QP8. This central object tends to describe the general view with respect to the subject therefore questionnaire parameters in class 1 which happens to be the major class with 5 questionnaire parameters is centered on the financial implication of solar energy generation.
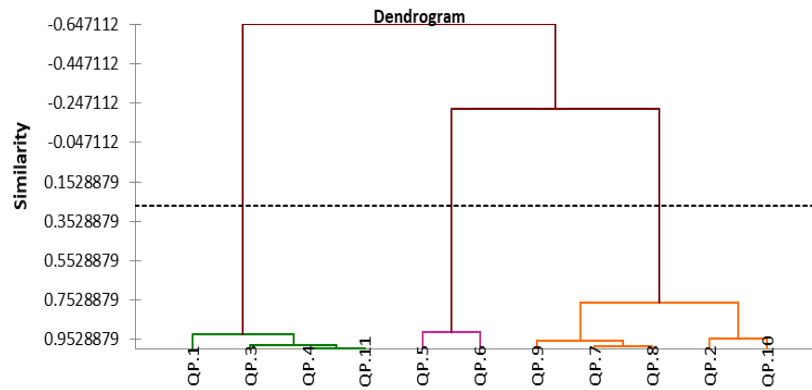
Figure 2 Dendrogram with respect to AHC output on Residential Occupants data set.

Table 1 Central objects with respect to the resultant classes – (Shop-Owners).

Class	YES	NO	UNDECIDED
1 (QP.5)	106.0000	174.0000	125.0000
2 (QP.6)	91.0000	137.0000	177.0000
3 (QP.8)	138.0000	156.0000	111.0000

Table 2 Central objects with respect to the resultant classes – (Residential Occupants)

Class	YES	NO	UNDECIDED
1 (QP.3)	102.0000	45.0000	33.0000
2 (QP.9)	57.0000	60.0000	63.0000
3 (QP.5)	79.0000	77.0000	24.0000

With respect to the data set for residential occupants, from Figure 2, the responses of the sampled population as pertaining to the subject matter were clustered also into three major classes (class 1, 2 and 3) with questionnaire parameters QP1, QP3, QP4, and QP11 belonging to Class 1, while questionnaire parameters QP9, QP7, QP8, QP2 and QP10 belong to Class 2 and finally questionnaire parameters QP5 and QP6 belong to Class 3. It is also interesting to note that from Table 2, the central object in class 1 was QP3 while class 2 was QP9 and class 3 was QP5. The questionnaire parameters in class 2 which happens to be the major class with 5 questionnaire parameters are centered around the political influence of the government solar energy for energy generation as an effective energy management system.

In terms of the general awareness level of the study population with respect to the acceptance of solar energy as an effective management for energy, majority of the shop-owners are unaware (Figure 3); this is in contrast with that of the sampled residential occupants (Figure 4). The unawareness of majority of the sampled shop-owner could be attributed to their general education level as majority (86%) of the study population only have basic educational degree (Secondary School degree) of which 60% are the shop-owners while 26% are the sampled residential occupants.

In order to predict the probability of general acceptance (variable 'yes') of solar energy as an effective energy management system, Equation 4 presents the resultant model fitting the responses (data sets of the sampled population of shop-owners and residential occupants). Table 3 presents the goodness of fit statistic with the likelihood function equal to 0 which according to Reed and Wu [20] shows that the resultant logistic model is a perfect fit with regards to the collected data set (Figure 3). It can be inferred from Table 4 which presents the probability analysis with the fitted model that the chance of acceptance of a solar energy management system is 1% at with 59.5% rejection from the study population. But with general improvement of the awareness of solar energy potential of the study population there is a

99% chance of acceptance with an unawareness level or say a rejection of 40.51% from the study population [21, 22, 23].
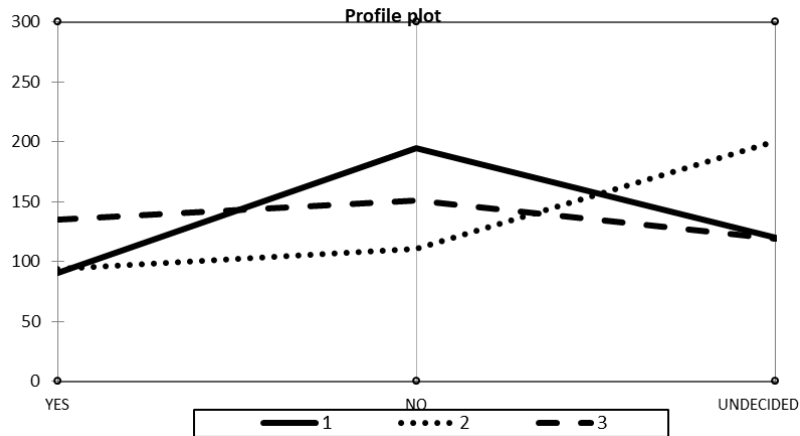

Figure 3 Profile plot for shop-owners with respect to awareness and acceptability of solar energy as alternate energy source.
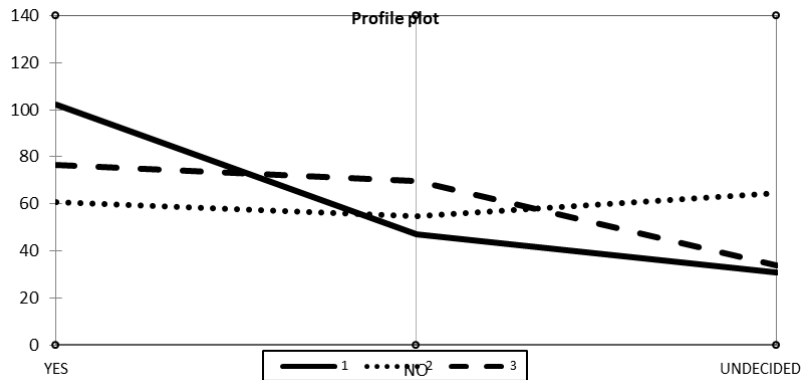

Figure 4 Profile plot for residential occupants with respect to awareness and acceptability of solar energy as alternate energy source.

Table 3 Goodness of fit statistics (Variable YES).

Statistic	Independent	Full
Observations	22	22
Sum of weights	22.0000	22.0000
DF	21	20
-2 Log(Likelihood)	30.4985	0.0000
R^2(McFadden)	0.0000	1.0000
R^2(Cox and Snell)	0.0000	0.7500
R^2(Nagelkerke)	0.0000	1.0000
AIC	32.4985	4.0000
SBC	33.5895	6.1821
Iterations	0	6

Table 4 Probability analysis with the fitted model (Variable YES).

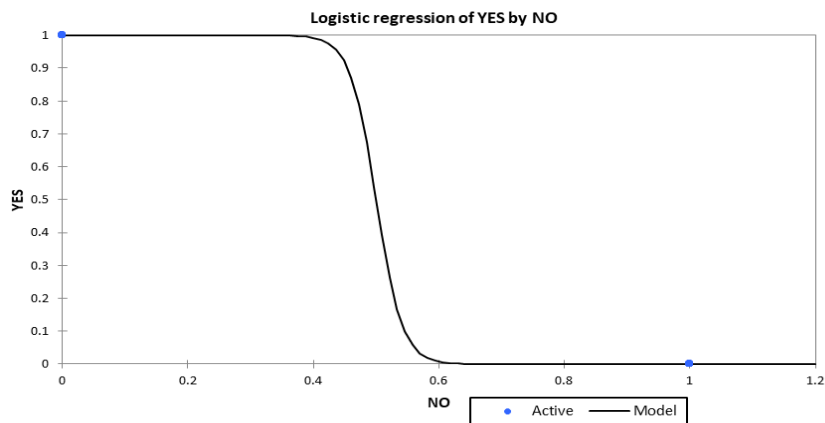| Probability (Solar Energy Acceptability)	NO
0.01	0.5949
0.05	0.5608
0.10	0.5454
0.20	0.5286
0.30	0.5175
0.40	0.5084
0.50	0.5000
0.60	0.4916
0.70	0.4825
0.80	0.4714
0.90	0.4546
0.95	0.4392
0.99	0.4051

Figure 5 Plot of the fitted model from logistic regression of 'Yes' by 'No'

$$\Pr ed\left(Yes\right) = \frac{1}{1 + e^{-(24.203 - 48.406No)}} \quad (4)$$

4. Conclusion

This study has been able to employ Agglomerative Hierarchy Clustering algorithm for assessing the acceptability and awareness of solar energy as an alternate energy source for effective energy management within Port Harcourt city. It has also developed a model for predicting the probability of acceptance of solar energy as an alternate source within Port Harcourt city applying the principle of logistic regression analysis. The resultant major cluster on application of Agglomerative Hierarchy Clustering on the data set of shop-owners was centered around the financial implication of solar energy generation while that of data set for residential occupants was centered around the political influence of the government solar energy of energy generation as an effective energy management system. The resultant logistic model from the data sets has a goodness of fit with the likelihood function equal to 0. From the developed model with respect to this study, it could be inferred that the chance of acceptance of a solar energy management system is 1% at

with 59.5% rejection. While with general improvement of the awareness of solar energy potential of the study population there is a 99% chance of acceptance with an unawareness level or say a rejection of 40.51% from the study population.

Compliance with ethical standards

Acknowledgments

Disclosure of conflict of interest

There is no conflict of interest.

References

[1]   Chigbo AM. The need to embrace Renewable energy technologies in Nigeria. Nigeria Journal of Solar Energy. 2010; 21: 182-186.

[2]   Akinbami JFK. Renewable Energy Resources and Technologies in Nigeria: Present Situation, Future Prospects and Policy Framework. Mitigation and Adaptation Strategies for Global Change. Kluwer Academic Publishers: Holland, The Netherlands. 2001; 6: 155-181.

[3]   Federal Ministry of Power and Steel. Renewable Electricity Action Program (REAP). Published by the International Centre for Energy, Environment and Development. 2006.

[4]   Ozoegwu CG, Mgbemene CA, Ozor PA. The status of solar energy integration and policy in Nigeria, Renewable and Sustainable Energy Reviews. 2017.

[5]   Sambo AS. Renewable Energy Technology for National Development: Energy Crisis in Nigeria. The Way Forward." Proceedings of the Annual General Meeting of the Nigerian Society of Engineers. Bauchi branch, Bauchi State, Nigeria. 2001.

[6]   Department of Energy & Climate Change, UK Renewable Energy Roadmap. 2011.

[7]   Mgbemene, CA, J Duffy, H Sun, SO Onyegegbu. Electricity generation from a compound parabolic concentrator coupled to a thermoelectric module, Journal of Solar Energy Engineering. 2010.

[8]   Sherwania AF, Usmani JA, Varun. Life cycle assessment of solar PV based electricity generation systems: A review Renewable and Sustainable Energy Reviews. 2010; 14(1): 540-544.

[9]   Tyagi VV, Nurul AA, Rahim NA, Selvaraj Jeyraj AL. Progress in solar PV technology: Research and achievement, Renewable and Sustainable Energy Reviews. 2013; 20(C): 443-461.

[10]   Sambo AS. Sustainable Energy Development in Nigeria. 2005.

[11]   Green MA. Solar Cells Operating Principles, Technology and Systems Application. Prentice Hall, Eaglewood Cliff, NJ, USA. 2005; 1982.

[12]   Nasir A. A Technology for Helping to Alleviate the Energy Problems; Solar Energy for Cooking and Power generation. paper presented at the 3rd annual Engineering Conference of Federal University of Technology, Minna, Nigeria. 2001.

[13]   Bradley D. Plants point the way to renewable energy, New Sci. 2005; 121: 8.

[14]   Rochell MS. The History of Solar Power Today. 2010.

[15]   Atukomi ZL, Oduola MK. Assessment of Solar Energy Acceptability Using Kolmogorov-Smirnov Distribution Test and Single Factor One-Way Analysis of Variance. International Journal of Renewable Energy and its Commercialization, 2021 (in press).

[16]   Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python. 2001.

[17]   Bererud, WA. Introduction to Logistic Regression Models with Worked Forestry Examples – Biometric Information Handbook. 1996; (7): 26.

[18]   Kleinbaum DG, Klein M. Logistic Regression, Statistics for Biology and Health -A Self-Learning Text. Third Edition. Springer New York Dordrecht Heidelberg London. 2010.

[19]   Alberti KGMM, Zimmet P, Shaw J. Metabolic Syndrome—a new world-wide definition. A Consensus Statement from the International Diabetes Federation Diabetic Medicine. 2006; 23: 469–480.

[20]   Reed P, Wu Y. Logistic Regression for Risk Factor Modelling in Stuttering Research. Journal of Fluency Disorders. 2013; 38: 88 -101.

[21]   Brounen D, Kok N. On the economics of energy labels in the housing market. Journal of Environmental Economics and Management. 2011; 62(2): 166–179.

[22]   Zorabi M. Mixed Method Research: Instruments, Validity, Reliability and Reporting Findings. Theory and Practice in Language Studies, Academy Publisher. 2013; 3(2): 254-262.

[23]   Cornish R. Statistics – An Introduction to Sample Size Calculations. Mathematics Learning Support Centre. 2006: 1-5.