Customer segmentation for East African Microgrid Consumers

Brian Otieno[1], Nathaniel J. Williams[12], Patrick McSharry[134]

[1]Carnegie Mellon University Africa, Kigali, Rwanda
[2]Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[3]African Centre of Excellence in Data Science, University of Rwanda, Kigali, Rwanda
[4]Oxford Man Institute of Quantitative Finance, University of Oxford, Oxford, UK

Abstract—**Many countries in East Africa have low access rates to electricity and this impedes their economic growth. A possible solution that has been explored is the use of microgrids. However, for microgrid business models to succeed, appropriate customers need to be identified and connected in order to sustain the business. PowerGen Renewable Energy (PowerGen) is a leading alternative energy provider in East Africa. For this paper, we used data from 277 customers over a period of nine months collected from four of its microgrids in Tanzania. This paper seeks to segment customers to determine distinguishing characteristics of different groups that would be beneficial for PowerGen. Machine learning techniques are used to construct predictive models and classification techniques. Empirical evidence demonstrates that demographic surveys, usually undertaken in advance of setting up microgrids, offer a rich source of information about the likely future consumption patterns of customers. The predictive advantages of having historical load profiles are also explored.**

*Index Terms--***Customer segmentation; Microgrids; Off-grid; Rural Electrification; Solar Power.**

I. Introduction

Off-grid microgrids are small electricity networks that have the ability to operate independently [3]. Given that they are usually less expensive than connecting remote areas to the central grid [5], they could play an important role in increasing energy access to many people who lack access in many parts of the developing world [6]. In Africa, only seven sub-Saharan countries, out of 48 [7], have access rates exceeding 50% [1]. Tanzania, a country in the East Africa region, had an electrification rate of 41% in 2015, but plans to increase this figure to 50% by 2020 and 90% by 2035 [2]. For the case of Tanzania, microgrids could offer a possible solution. Companies that develop microgrids are becoming increasingly active in such developing countries where policy conditions are favorable. One such company that has been at the forefront in developing energy solutions in East Africa is PowerGen. They are a leading microgrid developer and operator that was established in 2011 [3].

PowerGen describes itself as focusing on building the power utility company of the future whose business model will allow Africa to leapfrog the aging power models and infrastructure of the more developed world [4]. As such, PowerGen uses microgrids to offer clean energy as a service where customers access electricity via a pay-as-you-go model. This payment model works well, especially in areas where the upfront costs for installing the systems are prohibitively expensive for many customers [4]. However, for this model to be more effective, PowerGen must address some challenges. First, PowerGen would benefit from a data-driven approach for identifying the best prospective customers to connect to the network. This is especially important for them to avoid connecting people who may not generate enough revenue to justify the cost of connecting them to the network. This effect, where customers stop topping up, though connected, represents a substantial financial loss and can be reduced through effective customer targeting. Second, PowerGen need to determine the energy a given site will demand to improve the design and financial sustainability of the power system.

In response to these challenges, this paper seeks to: (1) determine which demographic factors PowerGen should consider to identify the most promising new customers and (2) segment customers into various groups to determine whether a potential customer would be profitable for PowerGen based on the group in which the customer would be classified. PowerGen has installed over 200 microgrids across countries in East Africa, including Kenya, Tanzania, Zambia, Uganda, Rwanda, Mozambique, South Sudan, and Somalia [3]. However, this study will focus specifically on the data collected at their microgrid sites in Tanzania.

II. Process and data Description

A. Process Description

In order to make a decision about where to set up a microgrid, PowerGen needs to identify a site that would be sufficiently profitable for them to invest in a microgrid. To achieve this, PowerGen usually considers the following aspects [3]:

1. Determination of the best site by looking at factors such as the vibrancy of economic activity and areas that are far away from the central grid that are unlikely to be connected to the grid in the near future based on rural electrification plans.

2. A customer application process helps identify the most promising customers and also the design of the reticulation network to support these customers. The customers who are connected to the network are usually ones that are expected to have the highest consumption and reside close to the town center. The challenge therefore is how to forecast

978-1-5386-4163-7/18/$31.00 ©2018 IEEE

which customers are likely to consume relatively large amounts of electricity.

3. Metering and Payment Systems

The microgrids deployed by PowerGen utilize a pay-as-you-go business model which heavily relies on mobile payment platforms and smart metering technology. Each customer has an account with PowerGen and as long as their balance is positive, the meter allows for consumption of energy while continuously tracking a customer's balance. It is therefore a prerequisite that potential microgrids sites have access to a cellular data network.

The customer screening challenge described above is usually hard to determine due to the lack of historical consumption data relating to a potential customer. Generally, this also makes it difficult to forecast demand. To solve this problem, historical consumption data of existing customers were first used to determine their average consumption and load profiles. Various models are constructed to assess the potential of identifying promising potential customers using only their demographic variables and the historical behavioral data of other existing customers.

B. Data Description

Two specific categories of data were provided for this study: demographic and behavioral. The demographic data is collected during the customer application survey before a customer is connected to the microgrid. The variables collected include location, mobile network, connection type (e.g. business or home), income sources and levels, type of buildings and their ownership, current energy sources, appliances owned and desired, whether applicant has children attending school and the assets owned.

Behavioral data refers to information collected once a customer has already been connected to the grid. This includes hourly electricity consumption and top up payments.

For the demographic data, a total of 277 customers spread across four sites were surveyed to provide 41 variables. Although the customers had three connection types: home, business and home & business, this study focused on only two of those, home and business, to have binary classification labels. The table below shows customers in the various sites with their connection types.

TABLE 1 SITES DATA CHARACTERISTICS

Site	Home (%)	Business (%)	Home & Business (%)	No. of Customers
Site I	55.56	22.22	22.22	45
Site K	70	11.11	18.89	90
Site L	53.41	35.23	11.36	88
Site R	68.52	24.07	7.41	54

III. METHODS

Understanding microgrid electricity consumption patterns is important for evaluating microgrid business models and informing technical design [3]. Because most microgrid customers have never had access to electricity, it is not possible to directly measure demand for electricity prior to deployment [3]. Customers could be successfully targeted if it were possible to identify demographic variables that contain useful predictive information. Specific variables could be selected as features for models offering predictive information about the growth, average consumption and load profile. In addition, behavioral data will also be considered in the models in order to utilize the new data that will have been generated by the customer. This data could be used to further determine whether a customer, who has been on the network for some time, will become more profitable in the future.

A. Load profiles

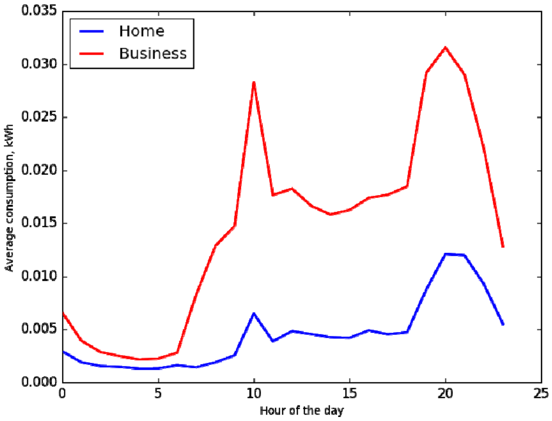

Figure 1: Hourly load profiles for home and business customers.

In order to derive the load profile for a given customer, the historical hourly consumption data was collected and the average consumption per hour was calculated over the course of a customer's time on the network. Thus, for each customer, this results in a load profile containing 24 data points i.e. 0 to 23 representing the 24 hours of the day. The customers were then grouped into business and home categories and their respective means taken leading to two representative load profiles, one for home and the other for business as shown in the Figure 1**.** As might be expected, the business load profile is larger than that of the home load profile and this difference is greatest during the afternoon.

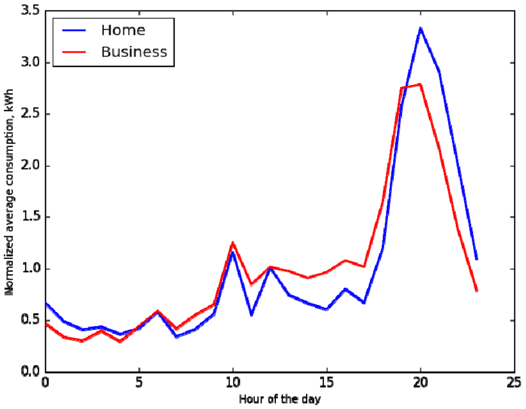

Figure 2: Normalized hourly load profiles for home and business customers.

It is also useful to derive a normalized load profile, whereby the variations throughout the day exhibited in the load profile are of interest rather than the actual level of consumption. This was achieved by taking the load profile for each customer and dividing by the mean hourly value, so that the derived quantities do not have any units and fluctuate about one. The resulting normalized load profiles for each customer were then averaged depending on whether their connection is home or business. The results are shown in the Figure 2. This analysis shows that the overall consumption patterns are remarkably similar apart from the business consumption being higher in the afternoon and home customers peaking slightly later in the evening.

B. Prediction

1) Linear Regression

This method estimates a linear mathematical relationship between features, also known as explanatory variables, and a specific dependent variable, which is usually assumed to be continuous [8]. The model parameters are usually estimated using the method of least squares [8]. In this analysis, the focus was to determine which variables are important in explaining the log of the hourly average consumption. Stepwise regression was used to identify the relevant variables. Later, an out-of-sample evaluation was performed using those variables with a 5-fold cross-validation approach.

C. Classification

1) Logistic Regression

Logistic regression is a method for predicting a binary dependent variable [8] i.e. predicting the likelihood of the dependent variable being equal to 1 given certain values of the features.

In this analysis, the challenge is to determine whether a connection type would be business-based or not. Initially, only demographic data will be considered, and later features based on behavioral data will be added. Stepwise regression was again used to determine the significant variables for constructing the model. Finally a logistic regression model is evaluated using a 5-fold cross-validation approach.

2) K-means

K-means is an algorithm that aims to partition N observations into k clusters so that the within-cluster sum of squared distance from cluster centroids is minimized [9]. In this application, the N observations refer to the PowerGen customers. It is also possible to compare the results from k-means where k is set equal to two with those from the logistic regression above. The objective was still to determine whether a connection type would be business or not given only demographic and later demographic and behavioral variables. The significant variables selected under stepwise regression were input into the k-means algorithm to obtain a pair of clusters.

IV. RESULTS

The models that were introduced in the previous section are now applied to the microgrid customer data. Results are grouped into two categories based on the features used to construct the models: (a) only demographic data; and (b) both demographic and behavioral data. Analysis of the demographic data alone relates to the potential of understanding customers that have not yet been connected whereas the combination of demographic and behavioral data could help to improve the predictability of existing customers.

A. Demographic Data

In order to determine which demographic features are most relevant, the analysis first focuses on selecting variables that contribute significantly to the classification of the connection type (Business or Home customers) and the prediction of the average hourly consumption. After selecting the features for each of these applications, a final model is constructed that facilitates the segmentation of the customers.

1) Classifying Connection Type

The challenge here is to correctly classify the connection type with home labeled zero and business as one. In this application, it was necessary to remove some demographic variables which could effectively lead to misrepresentation in the classification challenge. For example, variables identifying income sources (e.g. from business), connection type (home, business, home + business) were removed. The site variables were also removed because they would not be helpful when predicting what might be expected for a new site. In order to obtain a consistent data set across both the demographic and behavioral data, it was necessary to remove some customers with mean hourly consumption of zero, which resulted in a total of 236 customers. Customers with a mean hourly consumption of zero will lead to inconsistencies when developing the normalized load profile in Figure 2. For instance, to come up with the normalized consumption values for a customer, we divided all his/her consumption values with his/her mean. If the mean is zero, then we will have infinity as the result.

Correlation between each feature and the connection type is first calculated to inform about the relevance of each, whereby large absolute correlation values indicate that the feature contains predictive information. From Figure 3, the feature with the highest negative correlation is *building ownership own*. This could imply that people with their own houses tend to obtain connections for their homes. On the other hand, the variable with the highest positive correlation is *building ownership rent*, which implies that most businesses were probably on rented premises. Other positively correlated demographic variables include *using energy diesel*, *weekly electricity expenditure*, and *no school going children*. This could imply that business customers are more likely to use diesel energy. In addition, businesses tended to have higher expenditure on electricity (based on the survey) and generally no children attending school.

A logistic regression model was estimated using stepwise feature selection. In all of the results presented, the statistical significance of features is indicated by $p<0.1$ (*); $p<0.01$(**); and $p<0.001$(***). Table 2 describes the features selected and the performance of the classifier.

978-1-5386-4163-7/18/$31.00 ©2018 IEEE

TABLE 2: CLASSIFICATION RESULTS FOR DEMOGRAPHIC DATA.

Variable	Parameter estimates
buildingOwnershipOwn	-1.9104 ***
using energy diesel	1.3824 **
AUC (in-sample)	0.6962
Accuracy (in-sample, out-of-sample)	0.7754, 0.7205

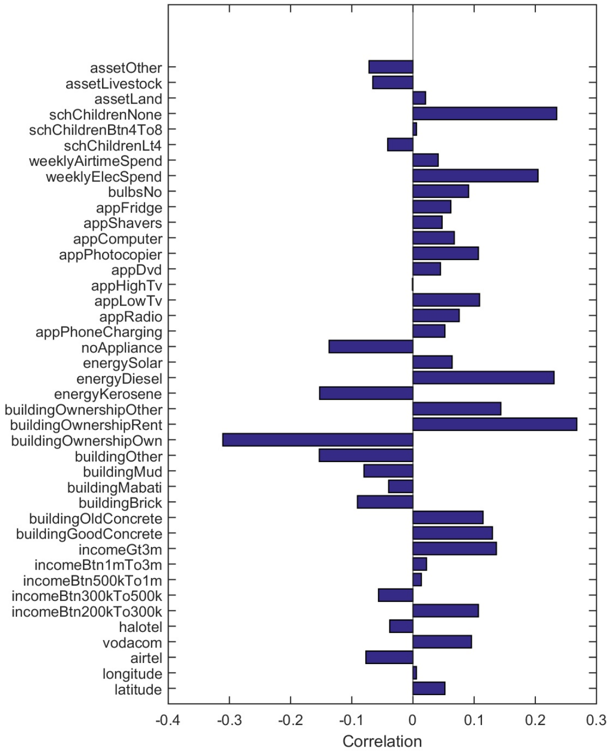

Figure 3: Correlation for demographic values and connection type.

The out-of-sample (OOS) accuracy was based on a 5-fold cross validation on logistic regression. A k-means model of the same data had an out-of-sample accuracy of 0.7627, indicating that k-means was a better classifier in this case.

2) *Predicting Average Hourly Consumption*

Another consideration which we deemed important was the average hourly consumption. The higher the average consumption of a customer, the more profitable that customer would be for PowerGen.

As in the logistic regression model above, correlation of all features with the dependent variable was first calculated in order to assess predictive potential. The dependent variable was log average hourly consumption. Furthermore, some additional features were added such as income sources and connection type, as they do not directly explain the dependent variable. The results are displayed in the Figure 4.

The highest correlations resulted from *income from business*, *weekly electricity expenditure*, *using energy diesel* and *Vodacom mobile number (we are considering dropping this because an area could be limited to a specific mobile network provider)*. This implies that customers who receive income from businesses are likely to be higher consumers of electricity. This seems to be supported by the highly correlated *weekly expenditure on electricity*. Customers who were using diesel as a source of energy generally become higher consumers once connected. Finally, it appears that business customers find Vodacom to be more reliable for them or rather home customers find Airtel to be more affordable. Generally, it seems that the more well-off people are, the more electricity they consume.

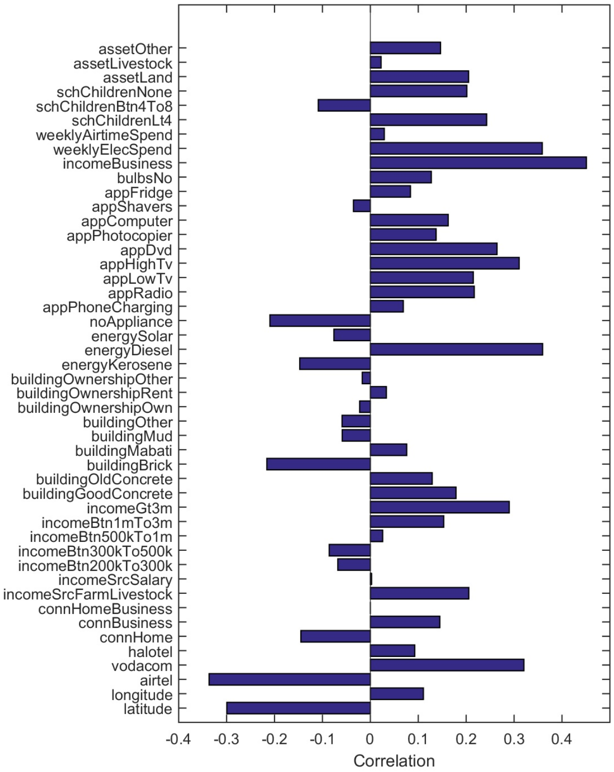

Figure 4: Correlation for demographic features and log average hourly consumption.

Stepwise feature selection identified a linear regression model for the log average hourly consumption (Table 3) where the OOS performance is based on a 5-fold cross-validation.

TABLE 3 REGRESSION RESULTS FOR DEMOGRAPHIC DATA

Variable	Parameter estimates
latitude	-0.19848 ***
Airtel mobile number	-0.41896 *
using energy diesel	1.3065 ***
owns high voltage TV	1.1135 ***
owns photocopier	3.3254 *
income from business	0.79501 ***
Adjusted R2 (in-sample, OOS)	0.3859, 0.3480

3) *Clustering*

All the features that were found to be significant in determining whether a customer had a business or home connection type and determining the average hourly consumption, were combined and fed into K-means to estimate a new model for clustering customers. In order to determine the number of clusters, a scree plot justifies selecting k=3 as shown in Figure 5.

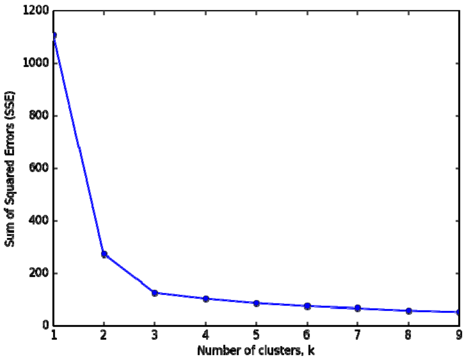

Figure 5 Scree plot for cluster on demographic variables.

In the results shown in Table 4, cluster 1 has medium business percentage and low consumption; 2 has high business percentage and high consumption whereas 3 has low business percentage, medium consumption. These results show that a high proportion of businesses does not always imply a high consumption level as indicated by cluster 1 which has a medium proportion of businesses but the least consumption. In addition, the results also demonstrate that it would be difficult to use demographic data alone to come up with clusters that do not contain businesses because a considerable fraction of businesses, greater than 10%, exists in each of the clusters.

TABLE 4 CLUSTERS AND THEIR CHARACTERISTICS

Cluster	Business customers (%)	Daily average (kW)
1	26	0.023089
2	37	0.319752
3	13	0.058254

B. Behavioural and Demogaphic Data

Behavioral data refers to digital metered records collected when a customer joins the network and can later be used to further classify a customer e.g. consumption and payment data. While PowerGen needs to identify which customers to select for connection, it is also important to monitor which customers have greatest potential in the future.

Features considered include the load profiles such as the hourly consumption at k hours and the normalized hourly consumption, kn. Another feature was constructed by adding the normalized hourly consumption between 09:00 and 17:00 to reflect the working day.

1) Classifying Connection Type

Logistic regression based on a stepwise feature selection approach was again employed to classify customer connection type. The analysis followed that of A.1 but with the behavioral features included and yielded an improvement in performance (Table 5). A k-means model of the same data had an OOS accuracy of 0.7627, suggesting that logistic regression was the better classifier in this case.

TABLE 5 CLASSIFICATION RESULTS FOR DEMOGRAPHIC AND BEHAVIORAL DATA

Variable	Parameter estimates
buildingOwnershipOwn	-2.2258 ***
19 hours	25.957 ***
15 normalized hours	1.3217 **
22 normalized hours	-0.61377 **
day 9 to17 hours normalized	-0.13835 *
AUC (in-sample)	0.8110
Accuracy (in-sample, OOS)	0.8136, 0.7888

2) Predicting Hourly Average Consumption

Similar to the linear regression in A2 for average hourly consumption using only demographic data, the analysis is repeated for both demographic and behavioral data. Features that would provide an unfair advantage were removed, e.g. hourly consumption data and some of its derivatives such as consumption during the day and night. Stepwise regression yielded a highly predictive model with a cross-validated R^2 of 0.6871. Customers with income from business and less than four children attending school were most likely to have high consumption levels.

TABLE 6 REGRESSION RESULTS FOR DEMOGRAPHIC AND BEHAVIORAL DATA

Variable	Parameter estimates
income from business	0.52498 ***
has school children less than 4	0.33879 *
1hours normalized	-0.72437 ***
6 hours normalized	-0.41129 ***
8 hours normalized	-0.3721 **
9 hours normalized	-0.69533 ***
14 hours normalized	-0.25213 *
19 hours normalized	-0.20536 ***
hourly normalized standard deviation	-1.5858 ***
hourly standard deviation	52.774 ***
Adjusted R2 (in-sample, OOS)	0.7344, 0.6871

C. Clustering

The variables chosen in B.1 and B.2 were combined and fed into K-means to estimate a new segmentation model for clustering customers. This model has the potential to distinguish between home and business customers and the level of consumption.

In Figure 6, the scree plot suggests k=3, and matches the previous number of clusters based on demographic data alone.

The results shown in Table 7 still demonstrate that it is difficult to find a cluster that does not have any business connection type as all clusters have a considerable percentage of businesses, at least 19%. On the other hand, the average consumption clearly displays three clusters with low, medium and high consumption levels. Cluster 2 is predominantly business customers and has high average consumption.

TABLE 7 CLUSTERS AND THEIR CHARACTERISTICS

Cluster	Business customers (%)	Daily average (kW)
1	19	0.047839
2	53	0.425236
3	24	0.229882

Table 8 provides a summary of the classification results. Adding behavioral variables increased the accuracy of some models significantly. For instance, OOS accuracy of logistic regression increased by 9.48% while OOS k-means remained the same.

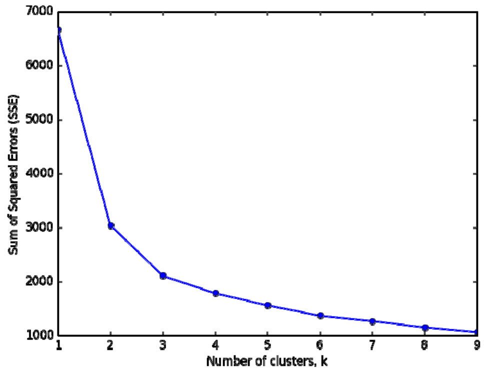

Figure 6 Scree plot for cluster on demographic and behavioral variables

TABLE 8 AGGREGATE RESULTS FOR CLASSIFICATION

Variables	Classifier	Accuracy	
		In-sample	OOS
Demographic	LR	0.7754	0.7205
	K-means		0.7627
Demographic + Behavioral	LR	0.8136	0.7888
	K-means		0.7627

Table 9 provides a summary of the results for predicting consumption levels. Adding behavioral variables increased the OOS R^2 value by 97.44%.

TABLE 9 AGGREGATE RESULTS FOR REGRESSION

Variables	Classifier	R Squared	
		In-sample	OOS
Demographic	Linear regression	0.3859	0.3480
Demographic + Behavioral	Linear regression	0.7344	0.6871

V. CONCLUSION

It is clear that demographic variables are useful in classifying whether a customer's connection type is business or home and also for predicting the average hourly consumption. Among the 236 customers analyzed, 172 were home and thus the classification benchmark is 0.7288 if every connection type were to be classified as home. K-means gave an OOS accuracy of 0.7627, clearly beating the benchmark. The demographic features that were important for classification were building ownership and use of diesel. The features that were important in predicting consumption were Airtel, using diesel, owning a TV, owning a photocopier and having income from business.

Addition of behavioral variables also significantly increased the classification and predictive power of most models. The OOS accuracy of logistic regression increased by 9.48% while the OOS R^2 value rose by 97.44%.

It also became apparent that businesses are interspersed within every cluster that was derived and thus it is very difficult to have a cluster without businesses. This implies that some homes and businesses could have almost the same consumption levels and probably similar consumption patterns. This further serves to prove that the connection type may not always influence how beneficial a consumer would be in terms of consumption. Finally, the load profiles were of similar shape for both home and business customers, although business was generally higher, especially in the afternoon.

For microgrid planners, the results suggest that a customer seeking a business connection may not always consume more than one seeking a home connection. For example, results from table 4 shows cluster 3 to be a better alternative to cluster 2 because of the higher daily average consumption. However, cluster 3 has lower percentage of business customers than cluster 2. As such, further research will be needed to find out which features determine a good potential customer within a specific connection type e.g. among the customers who want a business connection, which features will determine a good customer.

ACKNOWLEDGMENT

We would like to thank PowerGen Renewable Energy for providing us with the data to perform this research.

REFERENCES

[1] A. Castellano, A. Kendall, M. Nikomarov and T. Swemmer, "Brighter Africa: The growth potential of the sub-Saharan electricity sector", McKinsey & Company, 2015.

[2] United Republic of Tanzania, "POWER SYSTEM MASTER PLAN 2016 UPDATE", Ministry of Energy and Minerals,pp. 8, 2016.

[3] N. Williams, P. Jaramillo, B. Cornell, I. Lyons-Galante and E. Wynn, "Load Characteristics of East African Microgrids", PowerAfrica IEEE PES-IAS 2017, 2017.

[4] "The PowerGen story", PowerGen Renewable Energy, 2016. [Online]. Available: http://www.PowerGen-renewable-energy.com/the-PowerGen-story/.

[5] S. Szabó, K. Bódis, T. Huld, and M. Moner-Girona, "Energy solutions in rural Africa: mapping electrification costs of distributed solar and diesel generation versus grid extension," Environ. Res. Lett., vol. 6, p. 034002, Jul. 2011.

[6] International Energy Agency, "World Energy Outlook 2015," OECD Publishing, Paris, 2015.

[7] "The World Bank Group. Sub-Saharan Africa",. [Online]. Available: https://data.worldbank.org/region/sub-saharan-africa. Accessed Dec 13, 2017.

[8] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd ed. Toronto: John Wiley & Sons, Inc., 2000, pp. 1-10.

[9] J. Hartigan and W. A. Manchek, "Algorithm AS 136: A k-means clustering algorithm", Journal of the Royal Statistical Society, vol. 28, no. 1, pp. 100-108, 1979.