This is the accepted version of this article. The version of record is available at https://doi.org/10.1016/j.jobe.2021.102487

Evaluating the Determinants of Household Electricity Consumption Using Cluster Analysis

Eng L. Ofetotse[1*], Emmanuel A. Essah[2] and Runming Yao[2]

[1]School of Engineering and Environment, Kingston University, Kingston upon Thames, London, KT1 2EE, UK
[2]School of the Built Environment, University of Reading, Whiteknights, PO Box 219, Reading RG6 6AW, UK

*Corresponding author**:** e.ofetotse@kingston.ac.uk

Abstract

Identifying the determinants of household electricity use is a key element in facilitating the efficient use of energy. Even more so, segmenting households into well-resolved and characterised groups makes it possible to explore electricity use trends at more disaggregated levels, revealing consumption patterns and reduction opportunities for different consumer groups. Considering such groups, the drivers and implications of consumption trends can be better understood, bringing new insights into electricity use and offering opportunities to target policies and interventions that represent the needs of population sub-groups. For this reason, the aim of this research is to develop distinct household typologies using a k-means cluster analysis method. This was developed using questionnaire data of 310 households collected in a locality in Botswana. A feature selection procedure that maximises the silhouette was also developed to select the variables with the most significant clustering tendency. The analysis resulted in four distinct groups that are distinguishable by dwelling type, tenure, the number of rooms, the number of bedrooms, annualised electricity consumption and the number of appliances. The clusters identification enhanced the understanding of the fundamental factors underlying electricity consumption characteristics of different household segments. With this known, it is possible to identify those groups that offer the greatest energy saving potentials, thus providing insights for targeted demand-side management (DSM) and other possible strategies aimed at efficient energy use by customers.

Keywords

Cluster analysis; variable selection; k-means; silhouettes; cluster validity indices (CVIs); electricity consumption

Highlights

- A k-means clustering method and a feature selection method that maximises the silhouettes was developed
- Survey data from 310 households were utilised
- Four distinct groups that offer opportunities for different energy-saving strategies were discovered.
- Socio-economic factors were not significant in the discovery of clusters.
- Outcome of clusters provide insights that help support DSM strategies

1. Introduction

With the rise in residential electricity consumption and the need to balance demand and supply, a better understanding of residential energy demand characteristics is of vital importance. This is more so to assess and quantify the potentials and limitations (if any) of demand-side management (DSM) and other possible strategies aimed at encouraging efficient energy use by customers. The fact that there is diversity and variation in electricity consumption between different households makes it a significant barrier in understanding residential electricity consumption trends in greater detail. It also makes it harder to assess policy impacts at anything but highly aggregated levels, where most of the details are lost. This is especially considering the large number of customers, which makes it less feasible to design policies tailored to individual customers. For this reason, segmentation or customer grouping has come to the fore in the quest to provide distinguished groups without losing the important details in the data. This is supported by López et al. [1] who indicate that considering the large number of energy consumers, gaining correct knowledge of demand would require the grouping of customers with similar consumption habits. Räsänen et al. [2] also argue that providing information related to the consumption of similar customers in a neighbourhood makes electricity use information concrete and understandable. Iyer et al. [3] also assert that comparison of energy use with other similar customers may encourage energy-saving among customers hence also highlighting the need to group energy customers.

Typically, utility providers classify customers based on limited information such as the dwelling type and annual electricity consumption. Based on such classification, each customer is assigned a load estimate curve, which is used for billing and distribution management. However, as Räsänen et al. [4] indicated, it is not uncommon that changes in customers' way of life and electricity use do not mediate to the utility provider in a way that would allow them to update the load curve. Furthermore, previous research [5,6,7,8,9] has shown that residential energy consumption results from both the dwelling's technical characteristics and the composition of the household. Therefore, grouping households cannot be limited to just a few factors, such as energy consumption and dwelling type. Instead, all the factors that affect energy consumption should be incorporated into the grouping and those factors that show significant contribution to the creation of the groups be selected. In this way, changes in customer characteristics can be incorporated such that accurate groupings are carried out.

Customers' segmentation can be carried out through pattern recognition techniques that can be supervised or unsupervised [10, 11]. In unsupervised techniques, there exists a set of features, and the goal is to unravel the underlying patterns or groups. The most used algorithms under this category are cluster analysis techniques [11]. For supervised techniques, also known as classification [11], the groups are formulated from a set of labelled data. For example, discriminant analysis (DA), decision trees, support vector machines (SVM) which provides a statistical classification of samples contrary to the exploratory feature of cluster analysis [10, 11]. In the research presented in this paper, cluster analysis was used to create separate residential customer groups. The choice of clustering was for understanding which Wu [12] describes as to employ cluster analysis for automatically finding conceptually meaningful groups of objects that share common characteristics. This then helps in analysing,

describing and utilising the information hidden in the groups [12]. Therefore, cluster analysis is not about finding the right answer but about finding ways to look at a set of data that allows a better understanding.

This research aims to evaluate the fundamental factors underlying household electricity use of different household segments. With the different groups identified, it is possible to identify those groups that offer the most significant energy-saving potentials that can be used to optimise operations and better understand customers. At the same time, customers can use the analysis to understand their consumption profiles and behaviours to improve energy efficiency. The results of this study can also help utilities and policymakers develop energy efficiency strategies and policies, offer directed electricity-saving advice, improve energy consumption forecasting accuracy, and provide tailor-made energy services to specific customer segments. Furthermore, the energy consumption patterns of new customers can be estimated based on the designed segments.

2. Literature Review

Due to its effectiveness in data analysis and pattern recognition, cluster analysis has been widely used in various fields, including market research, web browsing, image indexing and community detection [12]. In residential energy consumption research, cluster analysis has been used to classify or group energy consumers [2,13,14,15,16], to identify habitual behaviour [17,18,19] and to detect household characteristics [20,21]. It has also been used to characterise and generate typical load profiles [4,22,23,24,25,26] as well as to detect outliers in data [27].

There are different clustering methods which typically differ from each other based on implementation, computational complexity and in the assumptions they make on the distribution of data [11]. In residential energy research, some studies have used hierarchical clustering [e.g. 4, 16, 20, 24] while others have used fuzzy c-means [e.g. 13, 15, 20]. Others have used k-means clustering [e.g. 13, 16, 23, 25, 26,28, 29, 30, 31,32]. Some researchers have applied multiple methods to the same data [e.g. 13, 16, 17, 20, 21]. In most cases, the k-means is used as a baseline and other methods compared to the baseline [21] while in other cases other methods are used to confirm the number of clusters while k-means is the main clustering method [17]. There are still inconclusive results regarding the best clustering method mostly because different methods produce different results on the same data. However, the most prevalent method in residential energy research is the k-means method. The advantages of the k-means method lie with its ease of interpretation, the simplicity of implementation, the speed of convergence and adaptability to sparse data [11, 33, 34]. k-means has been noted to have good cluster recovery qualities by Milligan [35] and Steinley [36] compared to other techniques hence its prevailed application. The k-means method's disadvantage lies with its sensitivity to noisy data and outliers [11, 33, 34]. A single outlier can dramatically increase the square error because such objects are far from the majority of the data. As a result, there have been modifications to the k-means method to counter its disadvantages as was carried out by Huang [37] and Ralambondrainy [38].

Most of the residential energy consumption studies use metered data either from smart meters [13,15,16,18,21,23,26] or field measurements [4,17,19,20,25] to discover the patterns and groups in electricity consumption profiles using different clustering techniques. In these studies, cluster analysis is carried out using load curves, and the characteristics of households such as socio-economic, building and appliance characteristics are then correlated with the developed clusters. In this case, households' different characteristics are not incorporated in the development of clusters and the most significant features selected. For example, Gianniou et al. [28] clustered energy consumers according to their consumption intensity before performing a correlation between energy consumption and the buildings and occupants' characteristics. The same procedure was employed by Yu et al. [19] who used cluster analysis and associate rule mining to relate clusters with user activities and appliance use. McLoughlin et al. [21] also created profile classes using clustering and then linked these to the household characteristics by applying a multi-nominal logistic regression. Rhodes et al. [23] clustered 103 homes using k-means clustering and proceeded to use a binomial probit regression method to determine if household characteristics could serve as predictors for clustering. Gouveia and Seixas [24] also clustered smart meter data sets and allocated survey data to the clusters. Although useful groupings are discovered using load curves and load profiles, the lack of incorporation of households' characteristics during the clustering process leaves a gap as far as assimilating changes in the households demographics is concerned. For this reason, in this paper, both energy consumption and the characteristics of households are incorporated in the segmentation of energy consumers and the variables that show the highest clustering tendency selected. A similar approach was applied by Wang et al. [39] and Viegas et al. [40] although they used smart meter and survey data for feature selection and extraction instead of survey data as used in this paper.

The limitation of this study lies with the availability of data. While challenges of obtaining household electricity consumption data seem to be a universal issue, the review of the literature has shown that the scale is greater in developing countries. Obtaining smart meter data in Botswana as with many other developing countries proved challenging. Also, current data on residential energy demand is not available, hence, there is a need that data is collected from scratch. The research presented in this paper, therefore, makes use of questionnaire data to discover distinguishing features of surveyed households and discover typologies that exist within a locality (neighbourhood).

In summary, this paper contributes to the current state of the art by presenting a cluster analysis of Botswana households electricity consumption. The paper also makes use of fourteen variables in developing a feature selection method that maximises the silhouettes. An extension of the concept of cross-validation to unsupervised learning by employing cluster validation indices resulting in variability estimates of the resulting clustering performance is also provided. There is limited research directed at exploring the consumption patterns and drivers of high energy usage more so in developing countries such as Botswana. The research presented in this paper has made a significant contribution to this area by investigating and characterising the drivers of energy use in Botswana households. However, there is still a need for further research in this area to contribute to knowledge. The information provided, therefore, forms the bases for further research in the area.

3. Research Design

One theme that seems to cut across all cluster analysis studies highlighted in the literature above is that cluster analysis is an ambiguous concept and that there is no one size fits all clustering technique. In this regard, cluster analysis requires domain knowledge as well as objective evaluation matrices. For this research, k-means, which is largely considered a partitioning method, was used. The choice of k-means was based on the need to discover distinct groups in the data. Furthermore, the k-means method has been noted to have good cluster recovery qualities compared to other techniques [35,36]. K-means is a centroid-based technique that assigns data into groups based on the distance between objects and a cluster centre point [11]. In this regard, given a dataset (D) containing w objects, the k-means method aims to partition these w objects into K clusters with two restraints: (1) The centre of each cluster is the mean position of all objects in that cluster. (2) Each object has been assigned to the cluster with the closest centre. The measure of the distance between the objects and the cluster centre points is measured using similarity measures. An objective function (squared error) is used to assess the clustering quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters [11].

Considering the ambiguity and difficulty researchers face when trying to utilise cluster analysis, it is essential that there are steps to follow while carrying it out to be able to obtain meaningful information out of the clusters. The cluster analysis steps (incorporating feature selection) utilised in this research have been illustrated using a flowchart (Fig. 1). Some aspects of the steps were extracted from the fundamental questions highlighted by Jain and Dubes [41] and later summarised by Jain [42].

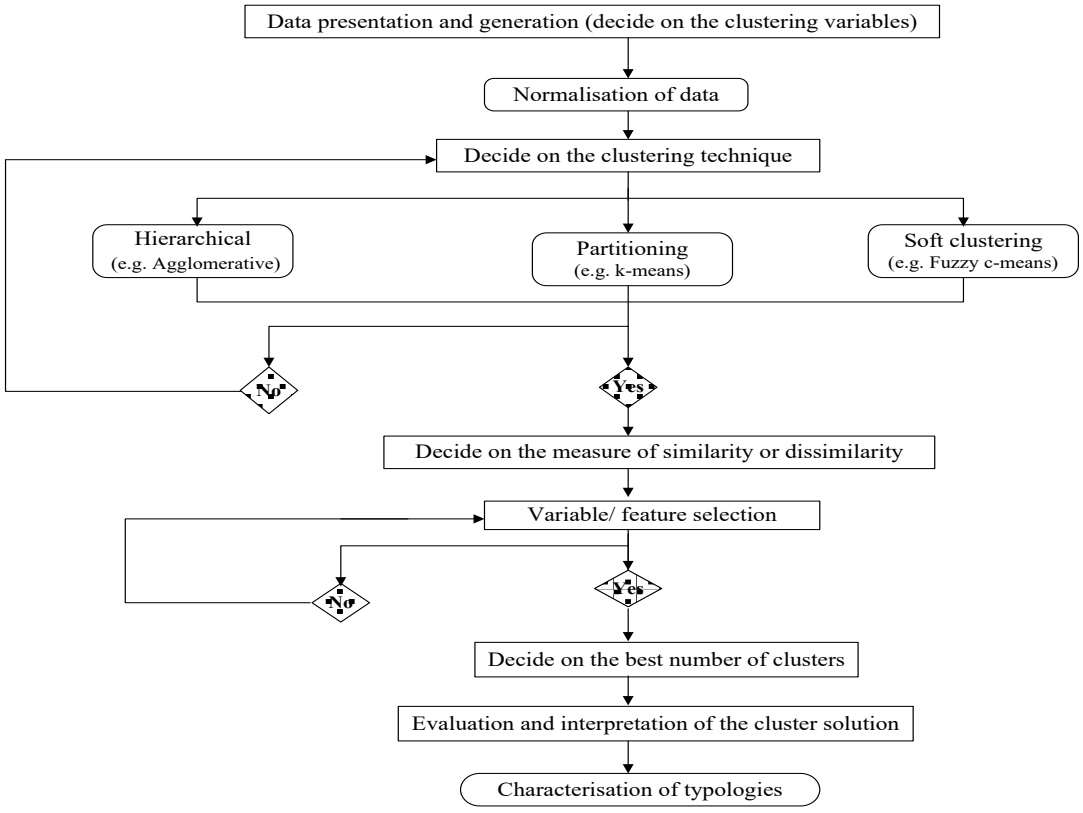

Fig. 1: A framework of clustering analysis incorporating feature selection and sequence of steps for analysis

Firstly, an exhaustive search was carried out whereby, having P number of features the possible feature subsets were 2^P-1. The generated subsets were then normalised using the min-max normalisation method. The choice of the min-max method was to preserves the inherent structure of the original data. This was because the aim of normalising the data was to rescale variables to the same scale without altering the data structure (restricting the variance). The mix-max method also helps to emphasise the importance of all the variables used for clustering. Furthermore, Steinley [43] examined the effect of different normalisation methods on the k-means clustering and found that the min-max method resulted in better clustering results compared to the use of z-scores and data that is not normalised. The normalisation also helps counter the disadvantages of k-means highlighted in Section 2.

The step that follows is the decision on the clustering technique. The other clustering techniques considered for this research were the hierarchy and soft clustering (fuzzy c-means) hence their inclusion in Fig. 1. However, the other clustering methods provided inconclusive results compared to k-means hence were not explored further. From the choice of clustering technique follows the selection of similarity measures. Similarity measures provide some measure that can determine whether two objects are similar or dissimilar. They also provide the concept of proximity and presents how close the objects are to each other or how far [44, 45]. The measures that can be applied to the k-means method include the Euclidean distance, Squared Euclidean distance, Manhattan (City block), and Chebyshev as reviewed by Shirkhorshidi [44]. For this research, the squared Euclidean distance was used. Compared to the other distance/similarity measures the squared Euclidean distance places progressively greater weight on objects that are further apart, which provides clearer separations between different clusters [45], which was the main objective of the clustering.

The step that follows is determining the best feature subset (see Section 3.4). Once the best feature subset is selected, the best number of clusters is determined. This is usually done by exploring a different number of clusters and determining the best one based on some criterion known as cluster validity index (CVI). CVIs are also used to assess the goodness of a clustering result by comparing them with other results from the same or different clustering techniques. The silhouette [46] CVI was used, although the Davies-Bouldin (DB) [47] and Calinski-Harabasz (CH) [48] were also explored yielding inconclusive results. After deciding on the best number of clusters, evaluation, interpretation and characterisation of the clusters can be carried out hence typologies created. The steps are further elaborated in Section 3.1- Section 3.3.

3.1. Data and Feature Generation/Search

The proposed methodology was applied to a dataset from 310 households. The data was obtained from a survey carried out in a locality (Block 7) in Gaborone, the capital city of Botswana (24.6282° S, 25.9231° E) over six months (June-December 2015). Block 7, highlighted in Fig 2, is one of the 120 localities in Gaborone with a population of 6197 people living in 2587 households. The locality comprises of privately and Botswana Housing Corporation (BHC) developed properties. BHC is a parastatal under the Ministry of Infrastructure and Housing Development that provides housing and building needs to government and local authorities. The properties in the locality can be broadly grouped into three categories, namely; standalone (making up 1,831 households), flats

(720) and townhouses (36) as shown in Fig 2. Altogether, BHC has developed 1823 properties in Block 7 within the three categories, including all 720 flats and 36 townhouses while the rest of the houses were privately developed.

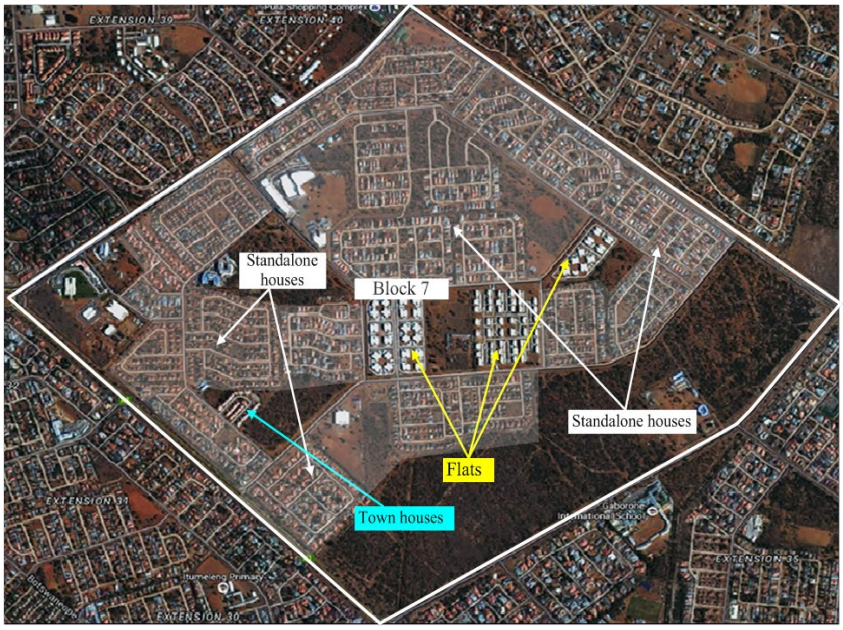

Fig. 2: Location view of the surveyed area showing the location of the different house types *Source*: Google Maps

Standalone dwellings have been categorised into low, medium and high-cost houses based on size and layout. Low-cost standalone dwellings as defined by the BHC have a total floor area ranging between 52 m^2 and 70 m^2 while medium-cost houses have a floor area ranging from 71 m^2 to 99 m^2. Any house with a floor area of ≥ 100 m^2 is considered a high-cost house. High-cost houses are usually identifiable by their sizes together with the presence of either single or double garage. In contrast, medium-cost houses have carports and low-cost houses are without either carport or garages. Townhouses are identical double-storey buildings that share partition walls. They range from two to three beds and a total floor area of 133 m^2 and 165 m^2. The townhouses are characterised by either an adjoining carport or garage. Furthermore, they are mostly designed for and occupied by government employees. Flats are high-rise multi-dwelling units ranging from one to three bedrooms. The number of units in one building range from six to as many as twenty-five and a floor area ranging between 67 m^2 and 93 m^2. Flats are considered low-cost housing mostly occupied by the low to middle-income occupants.

For this research, a stratified simple random sampling method was used to collect data. In this case, with a known number of the total houses in each category (flats, standalone and townhouses), measures were taken to obtain a representative sample of each category. This resulted in 360 questionnaires 50 of which were discarded due to incompletion resulting in 310 questionnaires remaining for analysis. Standalone houses encompass both detached and semi-detached houses. However, the semi-detached formed part of the discarded questionnaires hence in the study, detached houses instead of standalone will be used to describe this category houses. The 310 completed questionnaires encompassed 185 detached houses, 107 flats and 18 townhouses. These made up 10%, 15%, and

50% of the locality's total properties, respectively. This aligns with 10% of the population target usually deemed a sufficient representation of the population.

The survey data consisted of information on the building characteristics such dwelling type (flats, detached, townhouse), tenure (owned, rented, institutional), size (floor area), number of bedrooms, and the number of rooms. Information on socio-economic factors such as the number of occupants, presence of adults and children, age, gender, occupation (pre-school, student, employed, retired, homemaker) and their occupancy patterns, the amount paid towards electricity (bills) and income brackets (low, middle, high) were also requested was also collected. Furthermore, occupant characteristics and appliance ownership and usage were incorporated to investigate ownership, usage levels, size, model brand, and make of different appliances. The data obtained comprised of 37 variables that could be grouped into four groups and 14 subgroups, as indicated in Fig. 3.

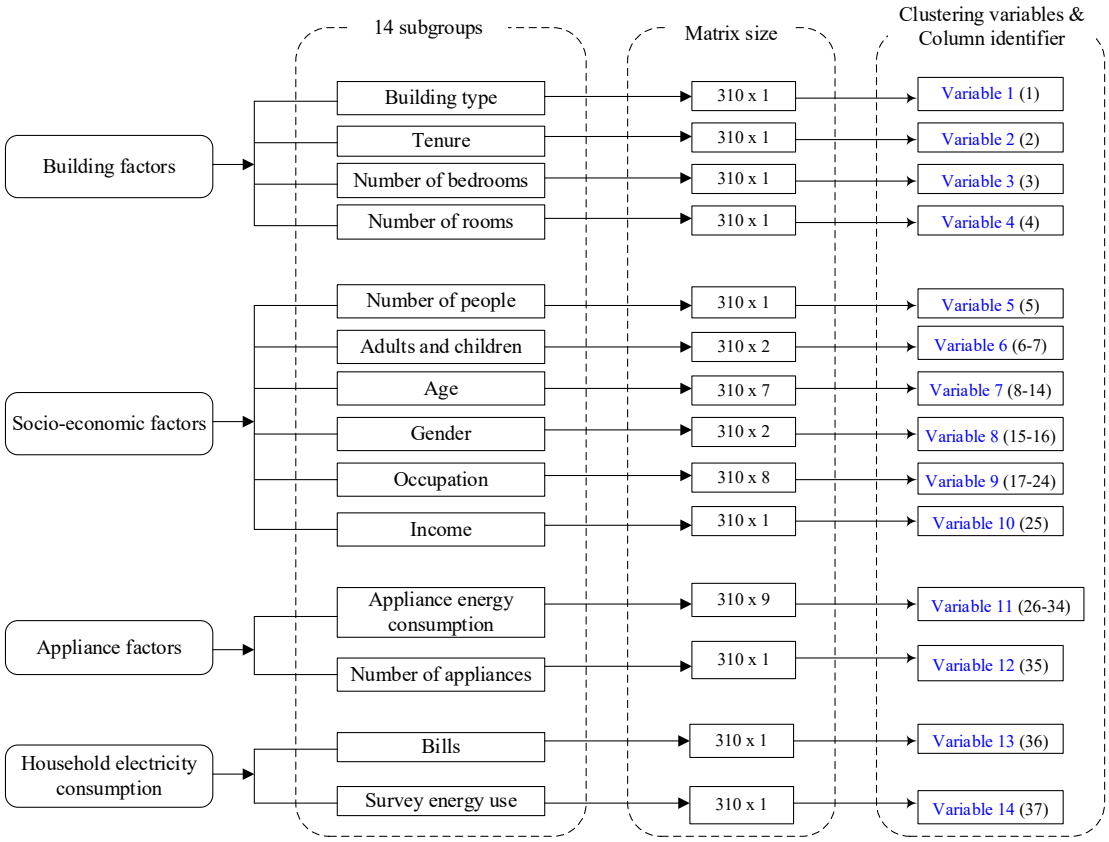

Fig. 3: Research data attributes and matrix sizes for clustering

In Fig. 3, some variables were in a single matrix while others were in multiple matrices. Adults and children were in two matrices, one for adults and one for children. Age was in seven matrices comprised of age groups 0-5, 6-18, 19-34, 35-49, 50-64, 65-79 and 80^{+}. Gender comprised of males and female matrices, occupation consisted of pre-school, students, part-time employments, fulltime employment, unemployed, homemaker/nanny, retired, other (including self-employed, infants). Appliance energy consumption was based on nine categories developed by Ofetotse et al. [49]. These were entertainment, communication (ICT), utility room, personal care, kitchen, heating and cooling, lighting, outdoor and miscellaneous. Household electricity consumption was based on two methods: (1) how much households were paying towards electricity (bills) and (2) appliance energy usage to

include ownership, usage levels, size, model brand and make of different appliances termed survey energy use. Comparing the two methods provided an $R^2 = 0.9489$ indicated a strong correlation between the methods. In this regard, it can be established that what households indicated to be paying and the intensity of use of their appliances correlated with each other. Hence, both consumption values can be used to assess households' electricity consumption trends.

The data was used to identify relevant features and extract household typologies. The process was carried out using the Matlab R2016b platform. An exhaustive search of order O (2^P-1) where P was fourteen (fourteen subgroups) resulted in 16,383 subsets. Though deemed computationally expensive, a complete search was selected to ensure that the best subset was chosen and the best grouping is discovered. The 16,383 subsets were used for feature selection and discover the household typologies.

3.2. Cluster Evaluation

Cluster evaluation is carried out through the use of cluster validity indices (CVIs). Three indices were used and compared: Davies-Bouldin (DB) [47], Calinski-Harabasz (CH) [48] and Silhouette (SIL) [46]. The DB index is the average similarity between each cluster and its most similar one. The clusters should have the minimum possible similarity to each other. Consequently, the number of clusters that minimises DB is taken as the optimal number of clusters. CH index uses a ratio of a between cluster means and within-cluster sum of squares. The maximal achieved index value indicates the best clustering of the data. Silhouette is a measure of how close each point in one cluster is to points in the neighbouring clusters ranging from [-1, 1]. Coefficients near -1 indicate objects assigned to the wrong cluster while those near zero indicate points that are not distinctly in one cluster. Coefficients near +1 indicate objects that are very distant from neighbouring clusters hence better clustering. Lletí et al. [50] suggest that as a rule of thumb, a mean silhouette value less than 0.25 shows that there is no substantial structure in the clusters found. Comparing the three indices showed that silhouette performed better, giving clear plots and results hence was selected as the CVI of choice. The average silhouette was used to measure how tightly grouped all the data in the cluster are and compare the results for the different number of clusters. The silhouette was also used to validate the selected variables, as discussed in Section 3.4.

3.3. Determining the number of clusters

For the selection of the optimum number of clusters, no standard objective selection procedure exists. A different number of clusters are explored in most cases, and the best number of clusters is evaluated based on a selected CVI (SIL in this case). For this research, the number of clusters was selected iteratively by executing the method with the number of clusters between 2 and 4 and selecting the best number of clusters from the output. This was based on the general rule of thumb for cluster analysis specified by Covões and Hruschka [51] which states that for a given dataset with P number of features the best number of clusters K ranges between $K_{min}= 2$ and $K_{max}= P^{1/2}$ [51]. Therefore, based on the fourteen features highlighted in Fig. 3, the best number of clusters is between 2 and 4.

3.4. Variable selection procedure

With the existence of 16,383 subsets, it was observed that some of the subsets are not true representations of the data in that they only represent one characteristic out of the thirty-seven and one subgroup out of the fourteen. It was also observed that the more variables there are, the lower the silhouette value hence, the worse the clustering. The subset with all variables (37 characteristics) had a silhouette value of 0.35, whereas subsets with silhouette values of 0.6 and above had ≤ 13 variables, as indicated in Fig. 4. This, therefore, indicates that including all features in the clustering affects the cluster discovery. Hence a feature selection needed to be carried out from which the best number of clusters could be determined.

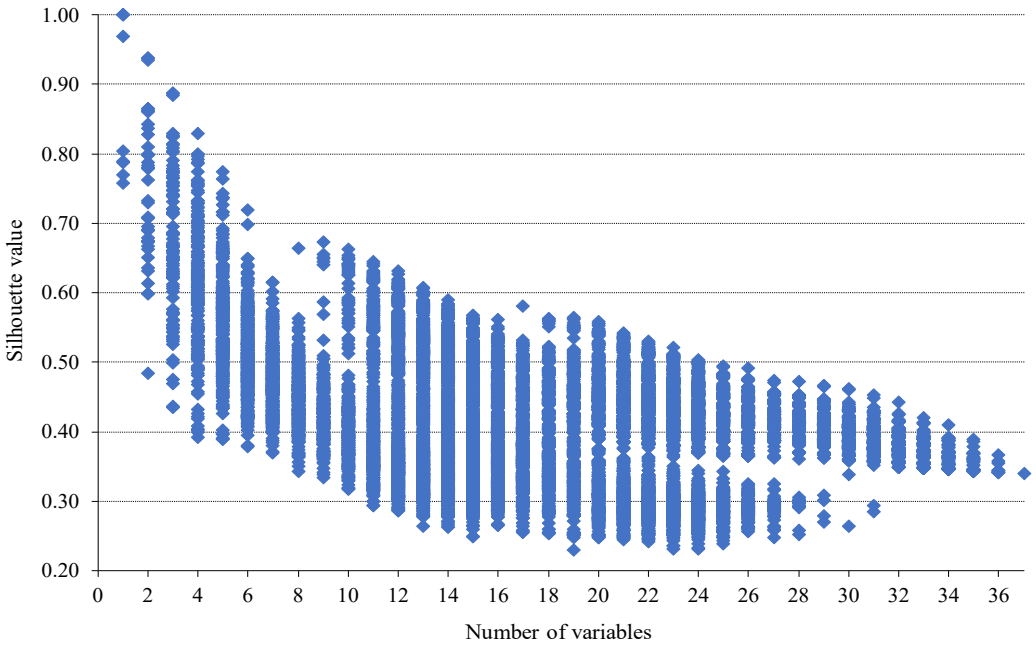

Fig. 4: Silhouette score against the number of variables

Feature selection selects subset $v \in V$ such that V is the 16,383 subsets and v is the optimum subset to be selected. Since this research used feature selection in clustering, it was sensible to use the clustering method in selecting features. Having generated the subsets after which they were clustered, a backward feature selection procedure was implemented to eliminate the less significant features one at a time based on their average silhouette scores. The initial procedure involved searching through the generated datasets to identify all the subsets containing each of the fourteen variables (see Fig. 3) and their varied silhouette values. Each variable was contained in 8,192 datasets of different combinations. The subsets' average silhouette values containing each of the fourteen variables were then calculated from which variable 7 was observed to show the least average silhouette (0.36) as indicated in Fig. 5 and Table 1. To confirm this, variable 7 (age of occupants) was added to the variables that showed the same average silhouette values one at a time. This started with variables 3, 4, 5,12,13,14 (average sil = 0.41) then 1, 2, 6 (average sil = 0.40) and 8, 10 (average sil = 0.39) all of which revealed that variable 7 had the lowest clustering tendency**.**

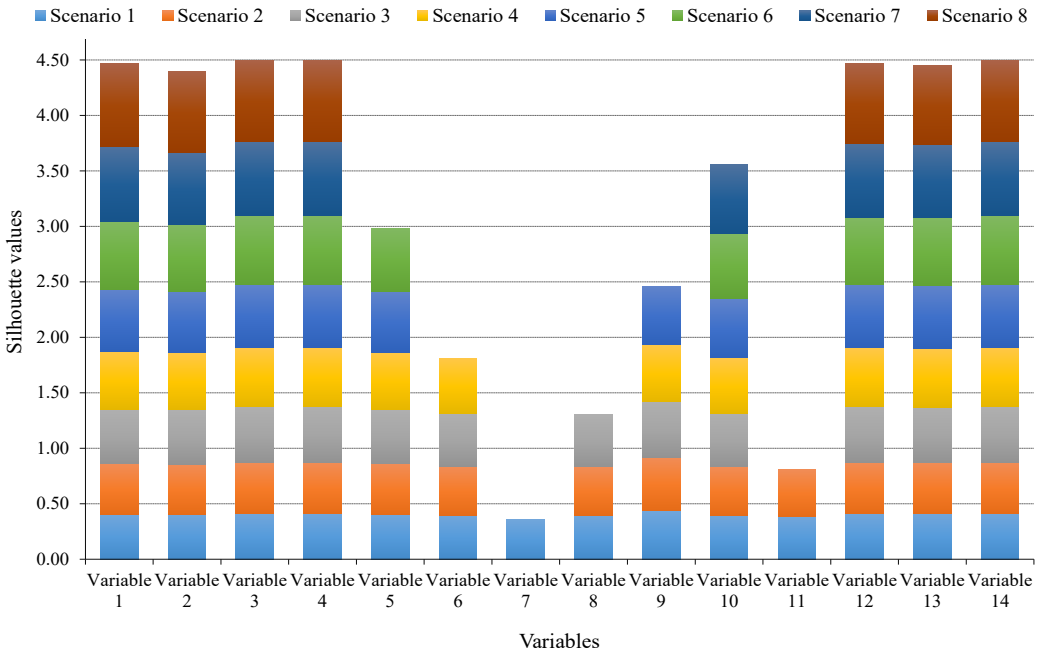

Fig. 5: Variables average silhouette values by scenario

Table 1: A summary of the variable selection procedure

Scenarios	Steps	Base feature sets	Least significant variable	Eliminated variable	Least Silhouette scores
1	1a	1-14	7	-	0.36
	1b	3,4,5, 12,13,14	7	-	0.50
	1c	1,2,6	7	-	0.56
	1d	8,10	7	7 (Age)	0.53
2	2a	1-6,8-14	11	-	0.42
	2b	3,4,12,13,14	11	-	0.54
	2c	1,2,5	11	-	0.62
	2d	6,8,10	11	11(Energy consumption of appliance groups)	0.50
3	3a	1-6,8-10,12-14	6,8,10	-	0.47
	3b	3,4,9,12,13,14	8	-	0.58
	3c	1,2,5	8	8 (Gender)	0.66
4	4a	1-6,9,10,12-14	6,10	-	0.50
	4b	3,4,12,13,14	6	-	0.59
	4c	1,2,5	6	6 (Adults and children)	0.68
5	5a	1-5,9,10,12-14	9	-	0.53
	5b	3,4,12,13,14	9	-	0.62
	5c	2,5	9	9 (Occupation)	0.71
6	6a	1-5,10,12-14	5	-	0.57
	6b	1,3,4,14	5	-	0.75
	6c	12,13	5	5 (Number of people)	0.70
7	7a	1-4,10,12-14	10	-	0.63
	7b	1,3,4,12,14	10	10 (Income)	0.75
8	8a	1-4,12-14	12,23,14	-	0.72
	8b	2,3,4	12	Inconclusive, feature selection stops	0.79

For scenario 2, the clustering was carried out without variable 7, generating 8,192 datasets. After searching through the generated datasets and obtaining the average silhouette values, variable 11 (appliance energy consumption) was noted to have the least value. This was also confirmed by adding variable 11 to variables 3, 4, 12, 13, 14; 1, 2, 5 and 6,8,10 from which it showed the least average value compared to the other variables. Based on this, variable 11 was eliminated. The same procedure was carried out for scenarios 3 to 8. Different variables were eliminated one at a time until eliminating features did not produce any better subset, and a sufficiently useful subset was obtained. This was therefore taken as the optimum subset. The final data set comprised of seven variables; dwelling type (variable 1), tenure (variable 2), the number of bedrooms (variable 3), the number of rooms (variable 4), the number of appliances (variable 12) billed annual energy consumption (variable 13) and survey annual energy consumption (variable 14). These variables were clustered, and the best number of clusters determined the results of which are discussed in Section 4.

4. Results and Discussion

The seven relevant features selected from Section 3.4, could be grouped into four clusters at best, as highlighted in Table 2. This was a departure from the rule of thumb stated in Section 3.3 regarding the number of clusters in relation to the number of features in that with seven variables K would have been between 2 and 3. However, four clusters exhibited the highest SIL hence took precedence over the rule of thumb. Also, increasing the number of clusters allows the identification of distinct patterns that would be useful to unravel and compare in order to create types of consumers for which different policy and energy reduction measures could be targeted. Each cluster can be considered as an archetype that categorises households in the surveyed locality (Block 7).

4.1. Characteristics of the Clusters

This section describes each cluster's underlying attributes and compares with the other clusters to better understand these archetypes' nature and composition. Table 2 summarises the clusters' underlying characteristics based on the seven classifying variables that would be further discussed below.

Table 2: Selected variables characterising the clusters

Characteristics		**Clusters**			
		Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Number of households per cluster		66	120	96	28
Variables	Dwelling type	Detached	102 Flats and 18 townhouses	91 Detached and 5 flats	Detached
	Tenure	Rented	Rented	Owned	Owned
	Average number of bedrooms	3	2	3	4
	Average number of rooms	9	7	9	11
	Bills average annual energy consumption (kWh)	6,300	5,000	6,100	11,100
	Survey average annual energy consumption (kWh)	6,300	5,000	6,200	11,200
	Average number of appliances	39	30	41	58

To provide more context, although socio-economic variables showed less clustering tendencies, they are presented in Table 3 to emphasise the differences and similarities between the discovered clusters.

Table 3: Socio-Economic Characteristics of the Clusters

Characteristics		**Clusters**											
		Cluster 1			**Cluster 2**			**Cluster 3**			**Cluster 4**		
		Number of Persons per Household											
		Min	Max	Ave	Min	Max	Ave	Min	Max	Ave	Min	Max	Ave
Household Composition	People	1	8	4	1	8	4	1	8	4	2	7	5
	Adults	1	7	3	1	6	2	1	6	3	2	5	3
	Children	0	5	1	0	5	2	0	4	1	0	3	2
	Males	0	5	2	0	4	2	0	4	2	0	4	2
	Females	0	6	2	0	6	2	0	5	2	1	5	3
		Number of Household with at least one member (%)											
Age of Household Members	0-5 yrs	45			32			36			46		
	6-18 yrs	50			57			57			86		
	19-34 yrs	85			73			84			82		
	35-49 yrs	67			77			74			23		
	50-64 yrs	11			10			23			11		
	65-79 yrs	0			1			2			4		
	80+	2			0			2			0		
Employment Status	Pre-school	38			28			29			36		
	Student	68			67			76			93		
	Part-time	9			8			5			14		
	Full time	98			100			99			93		
	Unemployed	9			8			18			11		
	Homemaker (maid)	38			20			30			54		
	Retired	3			3			7			4		
		Number of Households in Cluster (%)											
Average Monthly Income Range (Pula)	4,000-10,000	30			25			14			0		
	10,000-15,000	15			33			29			11		
	15,000-25,000	23			25			28			18		
	25,000+	32			17			29			71		

From Table 3, it was clear that there were subtle differences pertaining to socio-economic variables between clusters 1, 2 and 3, which further emphasises the low clustering tendency of the socio-economic variables. Also, the results of the survey indicated that the households in the locality had similar socio-economic characteristics. For example, there were no low-income (0-1,500 and 1500-4000 BWP) households from the survey. The households consisted mostly of the younger population (below 65 years) the majority of whom were in fulltime employment and/or students. This also coincides with the results of the census statistics CSO [52] that indicate that Gaborone comprises predominantly of the younger population. The cluster that stood out most was cluster 4 more so pertaining to the average number of people, the presence of age group 6-18, which coincides with the presence of students as well as the majority of households in the cluster being upper-high-income (25,000+BWP). From Table 2 and 3 it can be concluded that although socio-economic factors have been shown to have a significant contribution to electricity consumption, with Huebner et al. [9] and Huebner et al. [53] indicating that

they explain 24% and 22% of household electricity consumption respectively when it comes to grouping energy customers, they play a lesser role since they are more complex to quantify. In this case, household segmentation can be carried out even without the knowledge of the socio-economic characteristics of the households.

Cluster 1 (C1) is characterised solely by detached houses, all of which were rented. The average number of bedrooms was three, while the average number of rooms was 9, as indicated in Table 2 and Fig. 6. The annual electricity consumption from both bills and the survey averaged to 6,300 kWh. Fig. 7 and Fig. 8 show the range of energy consumption and appliance ownership of the cluster. The dwellings in this cluster ranged between the low-cost to medium-cost similar to cluster 3 hence the medium energy consumption of both clusters compared to clusters 2 and 4. While dwelling type has been used as a proxy for assessing the standard of living of household members in the absence of information on income levels [54], it can be noted that there are high-income earners living in low-cost and medium-cost housing. Therefore, although BHC classifies households in this locality based on their income (low, medium and high), this classification played a lesser role in the segmentation hence energy consumption trends. Therefore, other variables need to be considered when grouping households for energy characterisation. This is supported by Brounen et al. [6] and Huebner et al. [9] who indicate that the variation in residential electricity consumption is a combination of the dwelling's technical characteristics and energy-using habits.

Cluster 2 (C2) was the group with the highest number of dwellings (120). It comprised of all the surveyed townhouses (18) and 95% of the flats. All of the households were rented. The average number of bedrooms was two while the number of rooms averaged at 7. This cluster comprised properties that are mostly characterised as low-cost properties, although high-income earners occupied some properties. The annual electricity consumption from both bills and appliance survey ranged from 3,300 kWh to 12,300 kWh and averaged to 5, 000 kWh (see Table 2 and Fig. 7). Appliance ownership was least for this group, as indicated in Fig. 8, which explains the low energy consumption. The more appliances owned and used, the higher the energy consumption as highlighted by Bedir et al. [7] and Huebner et al. [53]. The range and mean of ownership of appliances was highest in detached houses (16-138; mean 43) compared to townhouses (27-50; mean 37) and flats (17-50; mean 29). Therefore, since cluster 2 comprises flats and townhouses, this explains the low appliance ownership and use.

Cluster 3 (C3) had most characteristics similar to that of cluster 1 except for the dwelling type and tenure. The cluster comprises detached houses and flats although flats made up only 5% of the dwellings. Cluster 3 comprised solely of owned houses, be it self-built or purchased while cluster 1 had rented houses only. The flats in cluster 3 were the only owned flats out of the 107 flats from the survey results. The similarity between cluster 1 and 3 was with regards to the number of bedrooms and rooms which averaged at 3 and 9 respectively for both clusters. The annual electricity consumption was also similar average wise although the range for cluster 3 was lower as depicted in Fig. 7. This could be explained by the small difference in the ownership and use of appliances by the households in the different clusters. Although cluster 1 had more households in the lower middle income (4,000-10,000 BWP), it still exhibited higher energy consumption in terms of range and average than cluster 3 which

was attributed to the types of appliances owned and frequency of their use as shown in Fig. 9 and Fig. 10. This supports arguments made above that households' energy consumption is a combination of both technical characteristics of the dwelling and the household's composition and background, including the household occupants and their electricity using habits.

Cluster 4 (C4) represented the least number of households (9%) from the survey. It was made up of the highest energy-consuming detached houses. All the houses were owned and had an average of four bedrooms and eleven rooms, making them the largest houses of all the clusters. The houses also possessed the highest number of appliances of all the clusters. These ranged from high-end entertainment appliances to swimming pool pumps and multiple air conditioning units as further discussed below. The high energy consumption of households in this cluster was attributed to larger dwelling sizes, possession and use of more appliances (see Fig. 8 and Fig. 9) and also high income (see Table 3) compared to the other clusters. The majority (71%) of the households had an average monthly income of 25,000+BWP. These combined factors have been attributed to high electricity consumption in other studies such as Baker and Rylatt, [14], Bedir et al. [7], Jones and Lomas [8] and Huebner et al. [9]. Therefore, although income did not contribute to the cluster discovery, its contribution to energy consumption cannot be ignored.

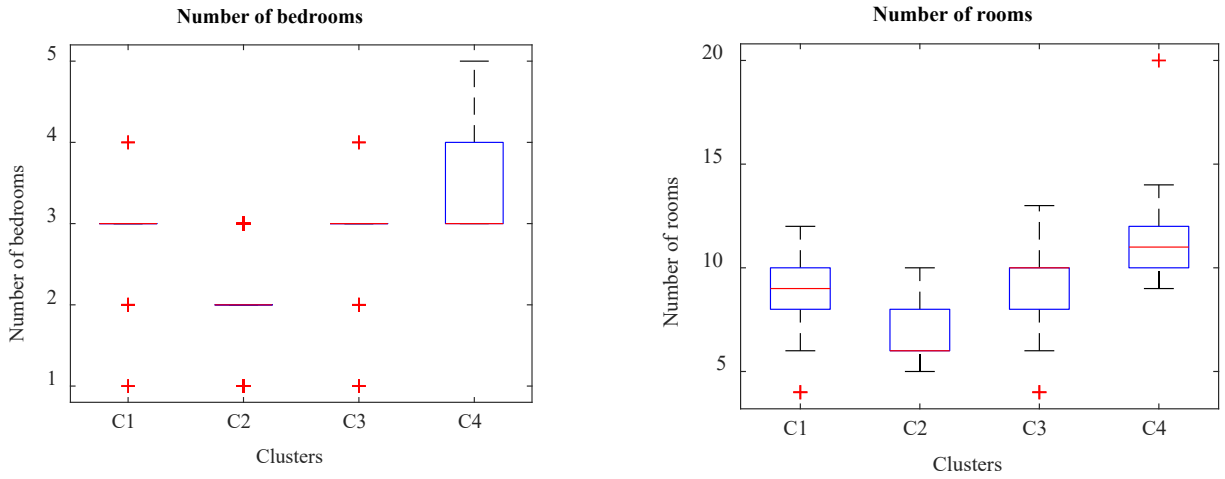

Fig. 6: Distribution of the number of bedrooms and rooms by cluster

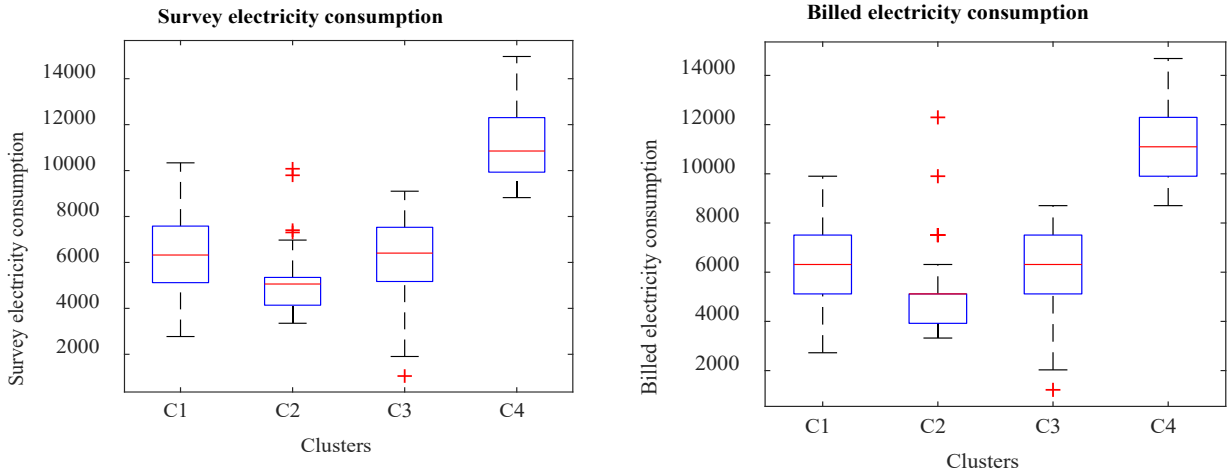

Fig. 7: Distribution of billed and survey annual electricity consumption by cluster

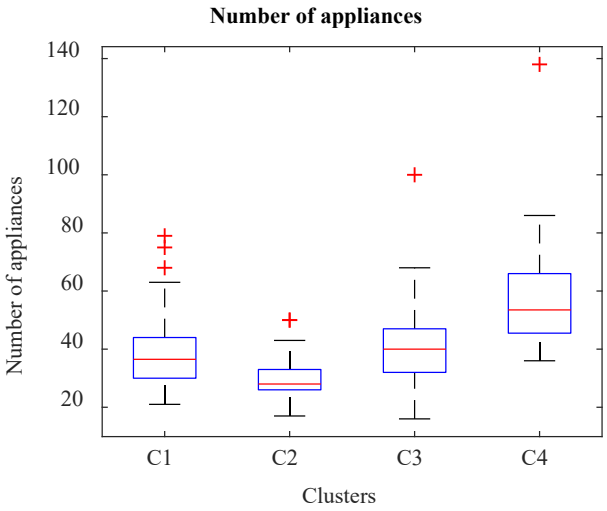

Fig. 8: Distribution of the number of appliances by cluster

In Fig. 9, the appliances owned by the different clusters is presented. Generally, in Botswana, ownership of appliances depends on usability and affordability. Therefore, there is no universal ownership of appliances hence low penetration rate of some appliances.

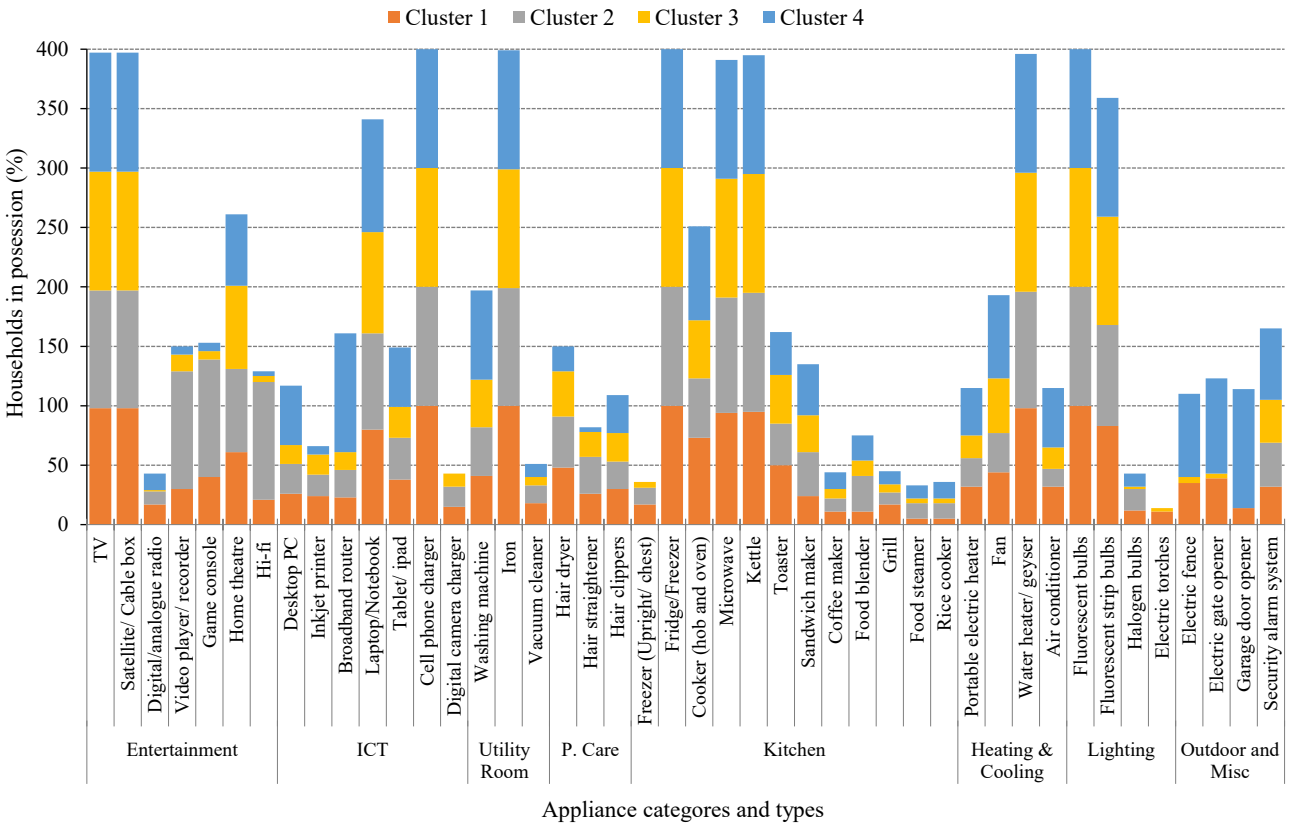

Fig. 9: Appliance penetration by cluster

The appliances indicated in Fig. 9 were only those appliances with significant ownership and energy consumption. Although there are no standard appliances in dwellings, the survey results revealed common appliances owned by households. These included appliances necessary for day to day living comforts such as cell phones, television sets, and fridge/freezers. As it would be expected, these were in more considerable possessions

in all clusters. However, cluster 4 owned more appliance in most cases, such as desktop computers, washing machines, electric cookers, air conditioning units and garage door openers. This, therefore, explains the high energy consumption of this cluster compared to others observed above.

To understand how appliance energy consumption differed between the clusters, Fig. 10 presents the clusters' energy consumption by appliance category. Heating and cooling made up the highest proportion of electricity consumption for all the clusters. This is mainly from water heaters that make up between 37%-41% of electricity consumption in the home, supporting BPC [55] findings where geysers were noted to be the highest contributors to electricity use in the home. Energy consumption attributed to heating and cooling was lowest for cluster 2 mainly because the cluster comprised predominantly of flats. Unlike detached houses and townhouses that use traditional storage tank geysers for water heating, flats use the kwikot Prisma or Ariston water heaters. The kwikot or Ariston water heaters differ from the storage tank geysers in that they heat water on demand, making them more water and electricity efficient. Varying capacities of the kwikot or Ariston water heaters are installed in kitchens and bathrooms of flats. Most of them have a power rating of 1,500 watts and are highly insulated compared to the tank storage water heaters, which are usually external to the building hence losing heat to the surroundings quickly.

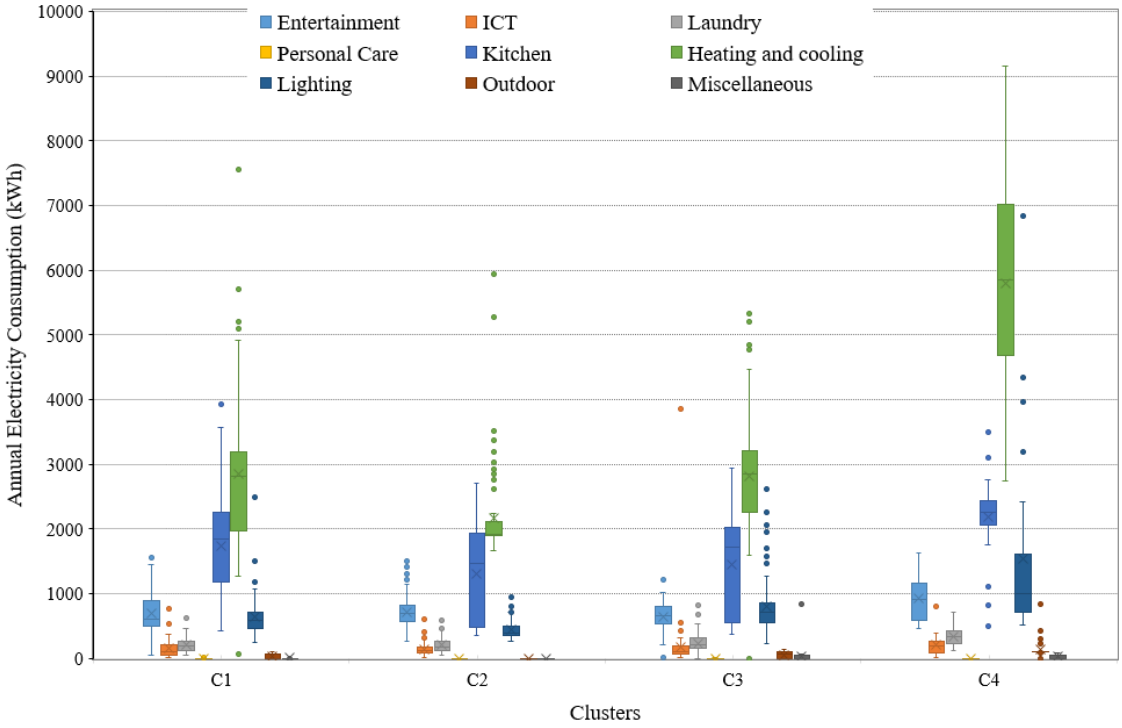

Fig 10: Energy Consumption of Clusters by Appliance Category

Air conditioning units were also significant contributors to energy consumption contributing up to 616 kWh on average per annum more so in detached houses. This was attributed to ownership and multiple (up to seven in some households) air conditioning units use. Kitchen appliances were the second-highest contributors to energy consumption mostly from fridge/freezers, cooker hobs and ovens, microwaves and kettles. These were highly owned in most households, as evidenced in Fig. 10.

For cluster 3 and 4 lighting energy consumption was higher than entertainment while for cluster 1 and 2 the reverse was true. This was due to the high ownership and use of lighting appliances such as compact fluorescent lights (CFLs) in cluster 3 and 4 and high ownership and use of entertainment appliances such as television sets, video players and game consoles in cluster 1 and 2. The high energy consumption of CFL lights was also attributed to the location in which they are used in the house. Most CFL bulbs were used in living rooms, bedrooms and outside the house. Bedrooms and living rooms are occupied for longer periods of the day compared to the kitchen (where fluorescent strip bulbs are mostly used) where the lights are used when the occupants are preparing meals and then switched off. Also, the CFL lights that are external to the building were usually kept on through the night to provide security hence contributing to energy consumption. Contribution of other appliances such as personal care, outdoor and miscellaneous was very low in all clusters mainly because of the low ownership, low power rating and less frequent use of such appliances compared to entertainment, kitchen heating and cooling and lighting appliances. In particular, outdoor appliances were not owned in flats and townhouses, which were the households in cluster 2 hence the zero consumption of these appliances shown in Fig. 10. Flats are multi-storey buildings in one compound all sharing gates and fencing hence no outdoor appliances specific to each apartment unit. Townhouses are enclosed houses with no security walls or electric fences hence no outdoor appliances.

The analysis provided in this section suggests that energy consumers are characterised by dwelling attributes to include dwelling type, tenure and building size, as well as appliance, attributes to include ownership, the type, the number and intensity of use.

4.2. Implications of discovered clusters to policy and DSM strategies

The characterisation of the households, regarding building factors, appliance factors and annual electricity consumption highlights and explains the wide range of electricity consumption characteristics within consumers of the same locality. This illustrates consumer segmentation's relevance for policies and measures design and implementation, tailored to energy reduction.

Due to high energy requirements and low energy supply, Botswana through the Botswana Power Corporation (BPC) put together some Demand Side Management (DSM) strategies, including the load curtailment program. This is a programme in which customer loads in the main cities (Gaborone, Francistown, Selibe-Phikwe, Lobatse and Jwaneng) are remotely managed using smart meters with remote disconnect capabilities [56]. The program aimed to limit power demand of individual households to not more than 2.3 kW (single-phase) and 6.9 kW (three-phase) during peak hours (between 06:00-1000 and 18:00-22:00) [56]. This strategy has minimised load shedding since its implementation in May 2015. Also, it has made domestic energy users more conscious of their daily energy use. However, measures should be taken to consider the different consumer groups and their energy needs. For example, as observed from this research results, some customers use high energy levels and can afford to do so. However, they are subject to maximum power limitations during peak demand hours, scheduled load

shedding, and unplanned brown-outs due to a lack of power generation capacity, weak grid infrastructure, and inefficient appliances. Therefore, it would be beneficial to develop strategies to provide reliable and sustainable energy for all customers.

Furthermore, education and awareness could be geared towards the cluster 4 households (who use twice more energy than the other clusters) to encourage them to save energy. These may include energy efficiency measures such as appliance substitution and behavioural changes. Furthermore, a reward system (incentives) could be implemented for those consumers who use less energy, such as cluster 2, to encourage them to continue using less energy. Such financial incentives may include dynamic pricing through real-time pricing as opposed to flat tariff rates, as suggested by Allcott [57]. Through this strategy, households who use less energy throughout a day would benefit from energy savings of up to 2% of the total electricity expenditure. Also, understanding the drivers of energy consumption of each cluster would help inform the strategies required to address the different energy requirements.

All in all, there is a significant difference between groups of consumers within the same locality, with some consumers consuming the minimum, while others are consuming three times the average along the year, more so pertaining to heating and cooling, which require a well-balanced interplay of policies and measures. For some consumers, the ultimate goal of energy reduction coupling energy efficiency measures (i.e. equipment's substitutions) and behavioural changes might be the focus. In contrast, for some, the incentivisation of continued use of low energy or switching to more affordable and sustainable renewable technologies may be the focus.

5. Conclusions

This paper evaluated the determinants of household electricity consumption using cluster analysis. This was done by applying the k-means clustering method and a feature selection method that maximises the silhouettes to a locality survey dataset. The 310 households from the survey were found to cluster into four distinct groups based on the dwelling type, tenure, number of bedrooms, number of rooms, billed energy consumption, survey energy consumption and the number of appliances in households. While socio-economic factors have been acknowledged to contribute significantly to household electricity consumption in previous studies, their contribution to discovering the characterising groups of households in the locality was less significant. Therefore based on the clustering results, it can be concluded that household segmentation can be carried out even without the knowledge of the socio-economic characteristics of households. Considering the four clusters' characteristic features, it can be concluded that there are four types of energy consumers in the surveyed locality. These are low energy consumers, two types of medium energy consumers and high-energy consumers. There were significant differences between the energy consumers with the high energy consumers using up to three times the energy of low consumers per annum, especially for heating and cooling. This was attributed mainly to the ownership and use of appliances.

The clusters identified provide a potential to better understand the underlying electricity consumption characteristics provided by different household segments. In this way, it is possible to ensure that interventions (such as the demand-side management strategy already in place) will encompass as much of the locality population as possible, and certainly, those groups which offer the most significant potential for beneficial impact together with the building and appliance factors that underpin these potentials. It would also be beneficial to extrapolate such methodologies to other localities and eventually to cities to obtain city-wide energy customer groups.

Although the method presented in this paper has been applied in the context of Botswana, it can be applied to other parts of the world using similar data. The proposed techniques can easily be implemented on large datasets hence the need to collect more data. The results can help utilities for better production-side management, such as developing new pricing policies geared at different customer segments and targeting specific customers to implement demand-side management solutions. Furthermore, the results can be a starting point in helping customers better understand their own consumption patterns and consumption behaviours to improve their energy efficiency.

Acknowledgement

The authors would like to appreciate the financial support from the Department of Tertiary Education and Financing at the Botswana Ministry of skills and development, without whom this work would have been possible.

References

[1] López, J.J., Aguado, J.A., Martín, F., Muñoz, F., Rodríguez, A. and Ruiz, J.E., 2011. Hopfield–K-Means clustering algorithm: A proposal for the segmentation of electricity customers. *Electric Power Systems Research,* 81(2), pp.716-724.
[2] Räsänen, T., Ruuskanen, J. and Kolehmainen, M., 2008. Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. *Applied Energy,* 85 (9), pp. 830-840.
[3] Iyer, M., Kempton, W. and Payne, C., 2006. Comparison groups on bills: Automated, personalized energy information. *Energy and Buildings,* 38 (8), pp. 988-996.
[4] Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K. and Kolehmainen, M., 2010. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy,* 87, pp. 3538-3545.
[5] Bartusch, C., Odlare, M., Wallin, F. and Wester, L., 2012. Exploring variance in residential electricity consumption: Household features and building properties. *Applied Energy,* 92, pp. 637-643.
[6] Brounen, D., Kok, N. and Quigley, J.M., 2012. Residential energy use and conservation: Economics and demographics. *European Economic Review*, 56(5), pp. 931-945.
[7] Bedir, M., Hasselaar, E. and Itard, L., 2013. Determinants of electricity consumption in Dutch dwellings. *Energy and Buildings,* 58, pp. 194-207.
[8] Jones, R.V. and Lomas, K.J., 2015. Determinants of high electrical energy demand in UK homes: Socio-economic and dwelling characteristics. *Energy and Buildings*, 101, pp. 24-34.
[9] Huebner, G.M., Hamilton, I., Chalabi, Z., Shipworth, D. and Oreszczyn, T., 2015. Explaining domestic energy consumption – The comparative contribution of building factors, socio-demographics, behaviours and attitudes. *Applied Energy,* 159, pp. 589-600.
[10] Theodoridis, S. and Koutroumbas, K., 2008. Pattern Recognition. 4th edition. USA: Academic Press
[11] Han, J, Kamber, M. and Pei, J., 2012. *Data mining: concepts and techniques*. Third edition. Boston. Morgan Kaufmann.

[12] Wu, J., 2012. *Advances in K-means Clustering: A Data Mining Thinking*. New York: Springer Publishing Company, Incorporated.
[13] Benítez, I., Díez, J.-L., Quijano, A. and Delgado, I., 2016. Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance. *Electric Power Systems Research,* 140, pp. 517-526.
[14] Baker, K.J. and Rylatt, R.M., 2008. Improving the prediction of UK domestic energy demand using annual consumption data. *Applied Energy*, 85 (6), pp. 475-482.
[15] Zhou, K., Yang, S. and Shao, Z., 2017. Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study. *Journal of Cleaner Production,* 141, pp. 900-908.
[16] Kwac, J., Flora, J. and Rajagopal, R., 2014. Household energy consumption segmentation using hourly data. *IEEE Trans Smart Grid*, 5 (1), pp. 420- 430.
[17] Abreu, J.M., Câmara Pereira, F. and Ferrão, P., 2012. Using pattern recognition to identify habitual behavior in residential electricity consumption. *Energy and Buildings,* 49, pp. 479-487.
[18] Haben, S., Singleton, C. and Grindrod, P., 2016. Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data. *IEEE Transactions on Smart Grid*, 7(1), pp.136-144.
[19] Yu, Z., Haghighat, F., Fung, B.C.M., Morofsky, E. and Yoshino, H., 2011. A methodology for identifying and improving occupant behavior in residential buildings. *Energy,* 36 (11), pp. 6596-6608.
[20] Gajowniczek, K. and Ząbkowski, T., 2015. Data Mining Techniques for Detecting Household Characteristics Based on Smart Meter Data. *Energies*, 8, pp. 7407–7427.
[21] McLoughlin, F., Duffy, A. and Conlon, M., 2015. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy,* 141, pp. 190-199.
[22] Labeeuw, W. and Deconinck, G., 2013. Residential Electrical Load Model Based on Mixture Model Clustering and Markov Models. *IEEE Transactions on Industrial Informatics*, 9(3), pp. 1561-1569.
[23] Rhodes, J.D., Cole, W.J., Upshaw, C.R., Edgar, T.F. and Webber, M.E., 2014. Clustering analysis of residential electricity demand profiles. *Applied Energy,* 135, pp. 461-471.
[24] Gouveia, J.P. and Seixas, J., 2016. Unravelling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy and Buildings,* 116, pp. 666-676.
[25] Do Carmo, C. M. R. and Christensen, T. H., 2016. Cluster analysis of residential heat load profiles and the role of technical and household characteristics. *Energy and Buildings,* 125, pp. 171-180.
[26] Al-Wakeel, A., Wu, J. and Jenkins, N., 2017. K-means based load estimation of domestic smart meter measurements. *Applied Energy*, 194, pp. 333-342.
[27] Khan, I., Capozzoli, A., Corgnati, S.P. and Cerquitelli, T., 2013. Fault Detection Analysis of Building Energy Consumption Using Data Mining Techniques. *Energy Procedia*, 42, pp. 557-566.
[28] Gianniou P., Liu X., Heller A., Nielsen P. S., Rode C., 2018. Clustering-based analysis for residential district heating data. *Energy Conversion and Management*, 165, pp. 840-850.
[29] Boudet, H. S., Flora, J. A. and Armel, K. C., 2016. Clustering household energy-saving behaviours by behavioural attribute. *Energy Policy,* 92, pp. 444-454.
[30] Fernandes, M. P., Viegas, J. L., Vieira, S. M. and Sousa, J. M. C., 2017. Segmentation of Residential Gas Consumers Using Clustering Analysis. *Energies,* 10 (12), 2047; https://doi.org/10.3390/en10122047
[31] Pan, S., Wang, X. R., Wei, Y. X., Zhang, X. X., Gal, C., Ren, G. Y., Yan, D., Shi, Y., Wu, J. S., Xia, L., XIE, J. C. and Liu, J. P., 2017. Cluster analysis for occupant-behavior based electricity load patterns in buildings: A case study in Shanghai residences. *Building Simulation,* 10 (6), pp. 889-898.
[32] Tureczek, A., Nielsen, S. P. and Madsen, H. 2018. Electricity Consumption Clustering Using Smart Meter Data. *Energies,* 11 (4), 859; https://doi.org/10.3390/en11040859.
[33] Rokach, L. and Maimon, O., 2005. Clustering methods. In: Maimon O. and Rokach L, Eds. *Data mining and knowledge discovery handbook*. Springer, Berlin, Chapter 15, pp. 321–352.
[34] Dhillon, I.S. and Modha, D.S., 2001. Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning,* 42 (1), pp. 143-175.
[35] Milligan, G.W., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika,* 45(3), pp. 325-342.
[36] Steinley, D., 2003. Local Optima in *K*-means clustering: What you don't know may hurt you. *Psychological Methods*, *8* (3), pp. 294–304.
[37] Huang, Z., 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery,* 2(3), pp. 283-304.
[38] Ralambondrainy, H., 1995. A conceptual version of the k-means algorithm. *Pattern Recognition Letters*, 16 (11), pp. 1147-1157.
[39] Wang, F., Li, K., Duić, N., Mi, Z., Hodge, B.-M., Shafie-Khah, M. and Catalão, J. P. S. 2018. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. *Energy Conversion and Management,* 171**,** pp. 839-854.
[40] Viegas, J.L., Vieira, S.M., Melício, R., Mendes, V.M.F. and Sousa, J.M.C., 2016. Classification of new electricity customers based on surveys and smart metering data. *Energy*, 107, pp. 804-81.
[41] Jain, A.K. and Dubes, R. C., 1988. *Algorithms for Clustering Data. Prentice-Hall advanced reference series*. Prentice-Hall, Inc., Upper Saddle River, NJ.

[42] Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters,* 31 (8), pp. 651-666.
[43] Steinley, D., 2004. Standardizing Variables in K-means Clustering. *In:* BANKS, D., MCMORRIS, F. R., ARABIE, P. and GAUL, W. (eds.) *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004.* Berlin, Heidelberg: Springer Berlin Heidelberg.
[44] Shirkhorshidi, A.S., Aghabozorgi, S., and Wah, T.Y., 2015. A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. *PLoS ONE*, **10** (12), e0144059. http://doi.org/10.1371/journal.pone.0144059.
[45] Hill T. and Lewicki P., 2006. Statistics: methods and applications: a comprehensive reference for science, industry, and data mining. United States: StatSoft, Inc.
[46] Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, pp. 53-65.
[47] Davies, D.L. and Bouldin, D.W., 1979. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2), pp. 224–227.
[48] Calinski, T. and Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics, 3 (1), pp. 1–27.
[49] Ofetotse, E.L., Essah, E.A. and Yao, R., 2015. Trends in domestic electricity consumption in Botswana. TMC Academic Journal, 9 (2). pp. 83-104.
[50] Lletí, R., Ortiz, M.C., Sarabia, L.A. and Sánchez, M.S., 2004. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 515(1), pp. 87-100.
[51] Covões, T.F. and Hruschka, E.R., 2011. Towards improving cluster-based feature selection with a simplified silhouette filter. *Information Sciences* 181(18), pp. 3766-3782.
[52] CSO, 2014. Population and housing census 2011 analytical report. http://www.statsbots.org.bw/sites/default/files/publications/Population%20%26%20Housing%20Census%20Dissemination%20analytical%20report%20.pdf (Accessed 15 April 2020).
[53] Huebner, G., Shipworth, D., Hamilton, I., Chalabi, Z. and Oreszczyn, T., 2016. Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes. Applied Energy, 177, pp.692-702.
[54] Yohanis, Y.G., Mondol, J.D., Wright, A. and Norton, B., 2008. Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic energy use. Energy and buildings, 40 (6), pp. 1053-1059.
[55] BPC, 2014**.** Saving electricity at home. https://www.bpc.bw/media-site/Pages/The-geyser.aspx (Accessed 12 April 2014).
[56] Ministry of Minerals Energy and Water Resources (MMEWR), 2015**.** Demand side management programme. http://www.gov.bw/en/Ministries--Authorities/Ministries/Ministry-of-Minerals-Energyand- Water-Resources-MMWER/MMEWR-Media/DEMAND-SIDE-MANAGEMENT-PROGRAMTO-BE IMPLEMENTED-FROM-TODAY-04-MAY-2015/ (Accessed 30 May 2015).
[57] Allcott, H., 2011. Rethinking real-time electricity pricing. *Resource and Energy Economics,* 33 (4)**,** pp. 820-842.