# Extracting Effective Features for Descriptive Analysis of Household Energy Consumption Using Smart Home Data

Hadiseh Moradi Sani, Soroush Omidvar Tehrani, Behshid Behkamal, and Haleh Amintoosi(✉)

Computer Engineering Department, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
{hadise.moradisani, omidvar}@mail.um.ac.ir, {behkamal, amintoosi}@um.ac.ir

**Abstract.** The household energy consumption has a large share of global energy consumption. To have better understanding of energy generation, management and surplus storage, we need to discover implicit patterns of consumers' behavior and identify the factors affecting their performance. The main goal of this paper is to descriptively analyze the pattern of household energy consumption using RECS2015 dataset. To this end, we focus on selecting the most effective subset of features from high dimensional dataset that leads to a better understanding of data, reducing computation time and improving prediction performance. The result of this study can help decision makers to investigate the living conditions of families in different levels of society to ensure that their life style is well enough or should be improved.

**Keywords:** Feature selection · Smart home · Household energy consumption · Correlation analysis

## 1 Introduction

With Rapid growth in world population, the development of urbanization and increased living needs, we are facing a sharp increase in energy consumption, specially the urban energy consumption which accounts for a large share of produced energy. Based on the reports [1], 31 Percent of U.S. households have problems paying their energy bills. Recent US Energy Information Administration (EIA) reports[1] also show that about one in three households reduces its basic needs, such as food and medicine, to pay its own energy costs, and 14 percent of them have received disconnection alerts due to not paying their bills [1]. Such issues and reports reflect the importance of monitoring energy consumption in the household sector, because applications such as demand-side management, energy management and demand-response management require an energy-monitoring. Another survey on energy consumption patterns of users in South Africa shows that citizen's income level has been identified to be an influential factor in

[1] https://www.eia.gov/consumption/residential/.


L. Grandinetti et al. (Eds.): TopHPC 2019, CCIS 891, pp. 240–252, 2019.
https://doi.org/10.1007/978-3-030-33495-6_18

energy consumption, and it has been determined that more than 70 percent of low-income families rely on energy sources rather than electricity. It also mentions that there exists a negative relationship between the electricity cost and its consumption [2].

Therefore, energy generation, management and surplus storage requires understanding consumers' behavior and the factors affecting it. Analysis of such factors and their impact on power consumption are among the important tasks in understanding the consumption patterns. Authors in [3] have investigated these factors and the role of each in urban residential energy consumption in China. Figure 1 categorizes these factors.

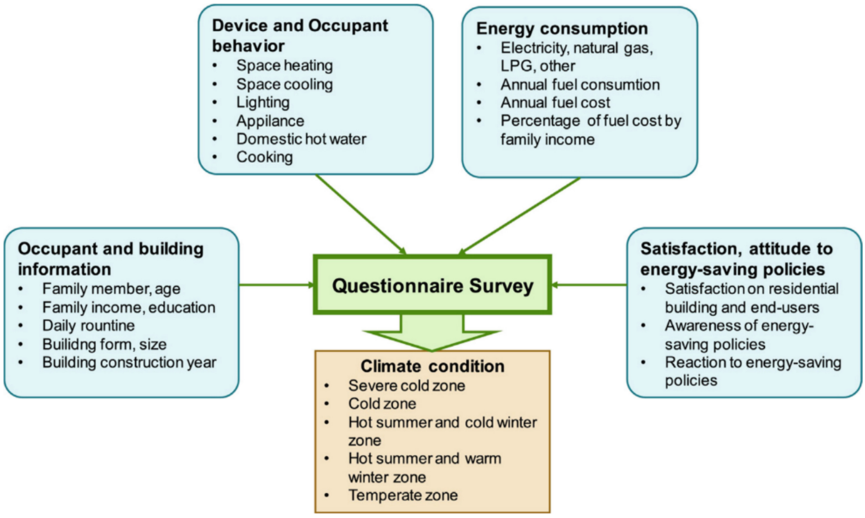


**Fig. 1.** Factors affecting the urban residential energy consumption [3]

As shown in Fig. 1, there is a wide range of factors affecting energy consumption. Thus, a main challenge for researchers in this field is selecting the most important factors in order to reduce the computational overhead. Lots of work has been done to increase the energy savings. Since the extracted data has high volume and diversity, big data techniques can be performed. Moreover, different data clustering algorithms, data analysis techniques and correlation between them have also been used in these works [4–7]. Other benefits of dimensionality reduction include avoiding over-fitting, resisting noise, and strengthening prediction performance in learning algorithms [8].

In this article, we consider RECS2015[2] dataset containing the data of residential energy consumption of US households in 2015, identify the most important features and remove non-relevant ones using feature selection techniques. Specifically, we first perform preprocessing with the aim of categorizing the features in a way to remove the unnecessary ones and select those features that are more relevant to the amount of

[2] https://www.eia.gov/consumption/residential/index.php.

energy consumed per year and the cost paid to the energy supplier. Once effective features are selected, clustering is done in two different ways on the data pertaining to these features in order to partition it to meaningful and accurate clusters. These clusters are then analyzed to infer useful information about the households' patterns of use, and its relation with their income and living spaces.

The structure of the paper is as follows. Section 2 discusses the related work. Section 3 presents proposed approach in three sub sections of introducing the dataset, preprocessing, and descriptive analysis. Section 4 presents a detailed analysis of the obtained results. Finally, Sect. 5 concludes the paper.

## 2 Related Work

In this section, we review the related works that present the effective features in energy consumption and discuss their feature selection methods. The work in [9] explores the relationship between energy consumption and its effective factors (focusing on the difference between rural and urban homes) in China, and shows that the most influential factor in energy consumption is the family income. Doing the similar research in the United States, the work in [1] examines the factors affecting energy consumption and their changes over time. The results show that the percentage of homes using only electric energy is increasing and the use of other fuels has declined.

Authors in [8] utilized ant colony optimization and the combination of the genetic algorithm (GA) and ACO (GA-ACO) to select the features in order to predict the short-term load forecasting based on the neural network. In this paper, effective factors on energy consumption are summarized in four categories: weather, time, economy, and random disturbances. The work in [10] evaluated the performance of four feature selection methods (autocorrelation, mutual information, RReliefF and correlation-based) and showed that all these methods are able to identify a portion of the highly-correlated features. Authors in [11] have used a two-step approach for choosing effective features in order to predict the cost of short-term electricity consumption, from the electricity market managers point of view. In the first step, using the Relief algorithm, the irrelevant candidate inputs are removed from the dataset, and then in the second step, using cross-correlation, redundant candidates are filtered. The selected features are then used in the forecast engine, implemented with the neural network. The work in [12] has used a filter-wrapper combination for feature selection with the aim of forecasting the short-term load. Taking the advantages of both filter and wrapper methods, this method first removes irrelevant and redundant properties based on the filter technique and then, by leveraging the wrapper technique and based on the Firefly algorithm, removes the extra features in a way not to decrease the accuracy. Also, in this paper, the need to use a different model to predict the energy consumption in special days has been emphasized.

## 3 Proposed Approach

The aim of our research is to select the most effective subset of features from high dimensional RECS2015 dataset of the household energy consumption. We believe that this leads to a better understanding of data, reducing computation time and improving prediction performance. There are three general categories for feature selection, namely, wrapper, embedded and filter methods [13]. Wrapper methods measure the "usefulness" of features based on the classifier performance. Naïve Bayes, SVM and K-means are some of the most famous examples of this group. Embedded methods are quite similar to wrapper methods since they are also used to optimize the objective function or performance of a learning algorithm. Decision tree algorithms such as CART and C4.5 are among the known examples of this category. In contrast, filter methods pick up the intrinsic properties of the features (i.e., the "relevance" of the features) measured via univariate/multivariate statistics instead of cross-validation performance. information gain, chi-square test, fisher score, correlation coefficient and variance threshold are well-known examples of filter methods.

In our research two attributes: 'the annual amount of energy consumption in thousand BTU' (known as TOTALBTU) and 'the total annual cost paid in dollars' (TOTALDOL) were considered as target attributes. The goal is to select the most effective features based on these target attributes. So, filter methods are known to be the best choice. To this end, we are taking the "relevance" of the features into account in order to analyze the relationship between features and target attributes. In the following, the dataset is described first. Then, two feature selection methods including Correlation Coefficient and Information Gain are explained and the results are compared. Finally, the results of descriptive analysis using clustering algorithms are presented.

### 3.1 Introduction of Dataset

The dataset investigated in this paper (RECS2015 (see footnote 2)) has been collected in 2015 by EIA [14] from American households and includes 740 attributes and 5,686 data records. Each data record represents a US household in 2015 based on data properties such as number and type of household appliances, patterns of energy consumption, structural characteristics of residential buildings, characteristics of household members and information of energy supplier. More detailed information is available on the EIA website [15].

### 3.2 Feature Selection

Generally raw data is normally vulnerable to noise, inconsistency and missing values, which may affect the result of data mining process. Hence, preprocessing is required to increase the data quality and the outcome of mining process. Data preprocessing includes four steps: data cleansing, data aggregation, data reduction and discretization. The first step is data clearing which is the process of detecting and removing inaccurate records within the dataset. The dataset used in this paper was first investigated for such records and it was observed that all records have valid values within their specified

range, no needing for cleaning. The second step is data aggregation, which is the process of combining data from different sources into a single coherent data store. This process had also been carried out by the EIA since the data collected from households and energy suppliers were all integrated in one single dataset. The major step of the pre-processing is the data reduction step, which is providing a reduced representation of dataset without compromising its integrity. Data reduction in normally considered as either reducing dimensions or reducing the number of records.

Since the dataset is highly dimensional containing 740 attributes, we need to apply an appropriate feature selection method to select the most effective features and reduce irrelevant dimensions. First, we analyze the correlation between all features and target attributes using Pearson's correlation method. This method measures the statistical relationship, or association between two quantitative variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. Then, we categorize features and analyze the most relevant feature in each category. We also use Information Gain as a metric to analyze the relevancy of categorized data. In the following, feature selection methods are expressed with more detail.

**Correlation Analysis on All Dimensions:** In this method, the correlation between every feature and the two target attributes (TOTALBTU and TOTALDOL) is calculated. This process is done for all 740 features and the features with highest correlation are going to be selected as the most important attributes. In should be noted that features expressing a single concept with several measurement units were removed.

Table 1 demonstrates these features and their correlation values with target attributes, ordered by the third column (labeled as 'Correlation with Class Label #1 TOTALDOL').

As shown in Table 2, the 12 features obtained by Correlation on Categorical Data method are among the features obtained when the correlation is computed for all features. This means that by categorizing the features first and selecting the most important attributes next, it is possible to identify and select the most important features out of 740 features.

**Correlation Analysis on Categorized Data:** In this step, we categorize dimensions of the dataset in order to select the most important features in each group. This process is performed with the help of experts. The result is partitioning 740 features into 8 categories, briefly described below.

1. *Home Characteristics*: In this category, reside the characteristics of the residential house such as: the geographical area, type of house (flat, apartment, etc.), number of floors, number of bedrooms, materials used in the construction of windows and ceilings, climate of that region, etc., constituting 47 features.
2. *Human Characteristics:* In this category, reside the characteristics of family members living in the house, such as: number of family members, number of members in/below the legal age, the level of education, annual income, number of days in a week they are at home, etc., There exist 20 features in this category.

**Table 1.** Selected features from all dimentions

| No. | Feature | ↓Corr. with TOTALDOL | Corr. with TOTALBTU |
|---|---|---|---|
| 1 | TOTALDOLSPH | 0.66 | 0.65 |
| 2 | TOTROOMS | 0.52 | 0.56 |
| 3 | WINDOWS | 0.52 | 0.55 |
| 4 | TOTALDOLNEC | 0.51 | 0.42 |
| 5 | TOTALBTUSPH | 0.51 | 0.85 |
| 6 | DOLELLGT | 0.47 | 0.38 |
| 7 | LGTOUTNUM | 0.47 | 0.48 |
| 8 | SWIMPOOL | 0.46 | 0.45 |
| 9 | DOLELCOL | 0.45 | 0.24 |
| 10 | TOTALBTUNEC | 0.44 | 0.42 |
| 11 | LGTINNUM | 0.44 | 0.44 |
| 12 | DOLELAHUCOL | 0.43 | 0.25 |
| 13 | TOTALDOLWTH | 0.43 | 0.14 |
| 14 | CELLAR | 0.42 | 0.54 |
| 15 | TOTALBTUWTH | 0.41 | 0.52 |
| 16 | SOLAR | 0.4 | 0.42 |
| 17 | STORIES | 0.4 | 0.48 |
| 18 | DRYUSE | 0.39 | 0.37 |
| 19 | WASHLOAD | 0.39 | 0.36 |
| 20 | DOLELAHUHEAT | 0.38 | 0.55 |
| 21 | MONEYPY | 0.37 | 0.34 |

**Table 2.** Selected features from categorized data

| No. | Category | Feature | Corr. with TOTALDOL | Corr. with TOTALBTU |
|---|---|---|---|---|
| 1 | Appliances Characteristics | LGTINNUM | 0.44 | 0.44 |
| 2 | | LGTOUTNUM | 0.47 | 0.48 |
| 3 | Energy Characteristics | TOTALBTUSPH | 0.51 | 0.85 |
| 4 | | TOTALBTUWTH | 0.41 | 0.52 |
| 5 | Behavioral Characteristics | WASHLOAD | 0.39 | 0.36 |
| 6 | | DRYUSE | 0.39 | 0.37 |
| 7 | Financial Billing | DOLELAHUHEAT | 0.38 | 0.55 |
| 8 | | TOTALDOLSPH | 0.66 | 0.65 |
| 9 | Fuel Consumption Profile | SOLAR | 0.4 | 0.42 |
| 10 | Home Characteristics | TOTROOMS | 0.52 | 0.56 |
| 11 | | WINDOWS | 0.52 | 0.55 |
| 12 | Human Characteristics | MONEYPY | 0.37 | 0.34 |

3. *Financial Billing:* In this category, reside the characteristics related to the bills paid for electricity consumption such as: the total amount paid, the amount paid for each electric appliance, etc. which constitute 57 features.
4. *Behavioral Characteristics:* This category includes those behavioral characteristics of family members which are related to energy consumption. These characteristics include: number of times household appliances such as stove, oven, microwave oven, washing machine, dishwasher, etc. are being used in a week, home temperatures in summer and winter, home temperatures when no one is at home, number of days cooling or heating appliances are being used, etc. 37 features are selected related to this category.
5. *Energy Consumption Characteristics:* In this category, features related to the way energy is consumed are located. These features define whether electricity has been used for cooling, heating or air conditioning, whether or not gas is being used for cooling, heating, cooking, etc., total amount of electricity consumption in 2015, the amount of electricity consumed by each appliance, and so on. The total number of feature residing in this category are 126 features.
6. *Fuel Consumption profile:* In this category, features related to the type and amount of fuel consumed within the house are located. They include the type of fuel being used to heat the bathroom, jacuzzi, and other household devices, the amount of solar energy used, and so on. 14 features related to this category are selected.
7. *Appliances Characteristics*: In this category, the characteristics of home appliances such as: number of fridges and their specifications (e.g., the year of production) and so on are described. The total number of features residing in this category is 114.
8. *Flag:* There are features with the role of flag for other features, that are considered to be in this category. Total number of flag features is 325.

Since the flag category did not provide useful information for data analysis, features belonging to this category were removed from the dataset. The remaining process was thus performed on features in categories 1 to 7.

Then, by performing the correlation function on each category, 12 features are identified with the highest correlation to two target attributes, as expressed in Table 2.

**Information Gain (IG):** One of the schemes for creating a decision tree is to use the notion of entropy. In this method, in each step, the attribute that mostly reduces the entropy is selected as the node of the tree (the root or the middle node). The reason behind this intuition is that attributes which reduce the entropy more are able to give more information about the data. In this section, the information gain method was performed on each of the 7 categories described above and in each category, features providing more information about the two target attributes are selected as the most important and effective ones. Table 3 shows these features and their information gain (IG) values for two target attributes.

Using information gain, 13 important features, most of which are similar to the features selected by two previous methods are identified that used the correlation criterion.

**Table 3.** Selected features by information gain

| No. | Category | Feature | IG with TOTALDOL | IG with TOTALBTU |
|---|---|---|---|---|
| 1 | behavioral characteristics | WASHLOAD | 0.12 | 0.11 |
| 2 | | DRYUSE | 0.12 | 0.11 |
| 3 | appliances characteristics | LGTOUTNUM | 0.16 | 0.17 |
| 4 | | H2OHEATAPT | 0.12 | 0.14 |
| 5 | energy characteristics | TOTALBTUSPH | 0.17 | 0.6 |
| 6 | | BTUELAHUHEAT | 0.13 | 0.34 |
| 7 | Financial billing | TOTALDOLSPH | 0.3 | 0.32 |
| 8 | | DOLELAHUHEAT | 0.14 | 0.31 |
| 9 | Fuel consumption profile | SOLAR | 0.12 | 0.14 |
| 10 | Home characteristics | WINDOWS | 0.18 | 0.21 |
| 11 | | TOTROOMS | 0.18 | 0.2 |
| 12 | Human characteristics | MONEYPY | 0.06 | 0.05 |
| 13 | | NHSLDMEM | 0.06 | 0.04 |

After applying three different methods for feature section, we understood that most of the features obtained from these methods are common. Thus, the intersection between the attributes selected by the above-mentioned three methods are 10 features as shown in Table 4.

**Table 4.** The feature suite selected by all methods

| No. | Feature | Mean | SD |
|---|---|---|---|
| 1 | WASHLOAD | 3.6 | 4.0 |
| 2 | DRYRUSE | 3.5 | 4.1 |
| 3 | LGTOUTNUM | 0.7 | 1.5 |
| 4 | TOTALBTUSPH | 33039.2 | 33492.6 |
| 5 | DOLELAHUHEAT | 24.2 | 37.5 |
| 6 | TOTALDOLSPH | 511.1 | 482.3 |
| 7 | SOLAR | −0.4 | 0.8 |
| 8 | TOTROOMS | 6.2 | 2.4 |
| 9 | WINDOWS | 35.8 | 11.3 |
| 10 | MONEYPY | 3.7 | 2.2 |

### 3.3 Descriptive Analysis

The result of preprocessing was the selection of 10 features which have the highest correlation with the target attributes. Since data records are un-labelled, unsupervised clustering schemes have been used to analyze the data.

Before going through the clustering process, we first investigated the information related to 10 selected features (expressed in Table 4). According to the data presented in Table 4, on average, the annual income of the households is $40,000 to $60,000. Houses are normally large with an average of 6 bedrooms. Solar energy is rarely used to generate electricity. The energy required to warm the living space is high, therefore, there is the possibility of large amounts of energy being wasted in heating/cooling the residential space. In the following section, the behavior of households in energy consumption is investigated using two clustering algorithms.

**Distance-Based Clustering:** The first clustering algorithm is *k-means* which is a distance-based clustering. Via this technique, data is divided into $k$ clusters in a way that the sum of the square of distances between the clusters is minimal [16]. *k-means* was first executed with different $k$ values to find the optimal value for $k$, i.e., the value that results in the optimal number of clusters. The result was $k = 4$. Then, *k-means* with $k = 4$ was performed to partition the data to four clusters. The size of clusters equal to 1541, 286, 899, and 2960 respectively. Figure 2.a represents a view of the data in each cluster.

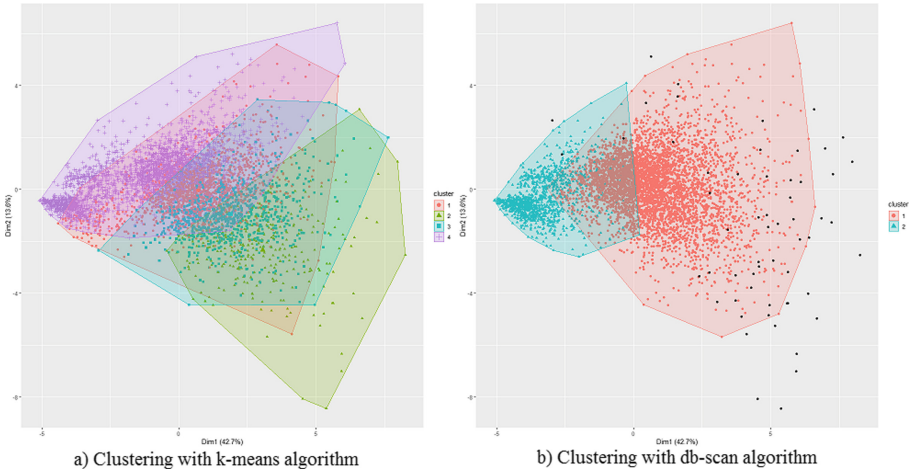


**Fig. 2.** Clustering result with k-means (a) and DB-SCAN (b)

Based on the results obtained, the data residing in the first cluster belongs to the families that frequently use the washing machine and the dryer during in a week and use a moderate number of bulbs. The households of this cluster use an average and even lower amount of energy to warm up their home space, so as to pay less for electricity. Solar panels are not used to generate energy. Also, these families have a relatively low annual income (generally between $20,000 and $60,000). This cluster includes 27.10% of the households.

Compared to the first cluster, families residing in the second cluster use higher amount of energy to warm the living spaces and water. So, they pay higher for electricity. Solar panels are not used in these households either. Households in this cluster have a high annual income (generally more than $140,000). Their houses are large and

have more rooms and windows compared to other three clusters. This cluster includes 5.02% of all households.

The characteristics of the data in the third cluster is a combination of the first two clusters. About 15.81% of households reside in this cluster.

Families residing in the fourth cluster generally do not have a washing machine and dryer, and the number of lamps used in their houses is very low. These households consume a very low amount of energy to heat their home space. The annual income of these households is very low (less than $4,000). This cluster contains 52.05% of households.

**Density-Based Clustering:** In density-based algorithms, the clustering is done based on the data density. In other words, the goal is to divide the dataset into subgroups of highly dense regions. Dense regions are then called clusters. We selected *DB-SCAN*[3], a density-based clustering algorithm which is also able to detect outliers -the data points that are in low density regions [16].

Running *DB-SCAN* on the dataset resulted in partitioning the data into two clusters, one with 4464 and the other with 1164 members. 58 records were also identified as outliers. Figure 2.b represents a view of the clusters made by *DB-SCAN*.

Analyzing the data of clusters show that families residing in the first cluster use washing machine and dryer. They also use a quite enough number of lamps in their houses. They also consume medium-to-high level of energy to warm their living spaces, resulting in a fairly high cost to be paid for electricity. The number of bedrooms and windows within the houses are quite large. Households in this cluster have an average-to-high annual income.

As for the second cluster, washing machines and dryers are not used. Moreover, no light bulbs are used outside the houses. The energy consumption is low, resulting in low cost to be paid. Houses in this cluster have a small number of bedrooms and windows. The annual income of families is very low.

Households in outlier data include those with very high or very low income. The amount of using washing machines and dryers in these two groups is very different, as they are used many times during the week. The amount of energy consumed in these households is also different. In general, households in this group show an unfamiliar behavior in energy consumption or their annual income differs from the rest of the community.

## 4 Discussion

In the previous section, two clustering algorithms were performed on the dataset. Here, the results obtained from these two methods are compared.

According to the results from *k-means* algorithm, low-income households (poor people), most of which are in cluster 4, generally do not use high electricity-consuming devices such as washing machines, and minimize the use of other electrical devices such as bulbs. This can be because they are either unable to pay the bills or they cannot

[3] Density-Based Spatial Clustering and Application with Noise.

use their own devices due to the debt to energy supplier. On the other hand, wealthy households which are generally in cluster 2, are not concerned about their energy consumption and their bills, because of their high income, Therefore, they consume energy without worrying about the cost. High energy loss (via windows for example) can also be observed in these households, which could be due to not caring about the cost to be paid.

Looking deeper at the results of *k-means* clustering, it is observed that *k-means* has not clustered the data in an effective way. In fact, only households in two clusters of 2 and 4 are different in their behaviors; the two other clusters, i.e., clusters 1 and 3, are a combination of both poor and wealthy households.

Next, we investigated the clusters provided by *DB-SCAN*. The obtained results are as follows:

- It is observed that low-income families are in cluster 2, accounting for about 20.4% of the total households. These households have small houses since the number of their windows and rooms are few. Therefore, in these houses, less energy is consumed for heating, and less energy is lost.
- Families with higher income reside in cluster 1, accounting for about 78.5% of the total households. They have big houses, causing energy to be wasted more.
- Outlier data is also about 1.02% of the data. Households residing in this cluster behave quite differently from the rest. So, putting them in a separate cluster as outlier prevents them from negatively affecting the behavioral analysis of ordinary households.

To conclude, results show that *DB-SCAN* clustering algorithm outperforms *k-means* in partitioning the data into more accurate clusters, as well as resulting valid pattern of the energy consumption.

## 5 Conclusion

The main purpose of this study is to analyze the pattern of household energy consumption using RECS2015 dataset on energy consumption which is collected from 5686 American households in 2015. Initially, a feature selection process was performed on all data to identify features that are more relevant to the amount of energy consumed per year and the cost paid to the energy supplier. By applying different filter methods 10 out of 740 features are selected. After dimensionality reduction, the dataset is clustered using two clustering algorithms of k-means and DB-SCAN in order to achieve accurate clusters and analyzing the life routine of households in terms of energy consumption. The results showed that density-based clustering provides more accurate and meaningful clusters compared to distance-based clustering.

It is worth mentioning that clustering helps to investigate the low-income community to know whether their living conditions (related to the energy sector) are well enough or should be improved. For example, a household may be forced not to consume the required energy due to severe financial problems or it may not be able to repair or purchase the necessary electric appliances. Similar results can also be concluded on high-income households to know whether or not they save energy even though they are not concerned about paying electricity bills.

## Appendix A. Description of the Features Mentioned in the Paper

| No. | Features | Description |
|---|---|---|
| 1 | BTUELAHUHEAT | Electricity usage for air handlers and boiler pumps used for heating, in thousand Btu at 2015 |
| 2 | CELLAR | Housing unit over a basement |
| 3 | DOLELAHUCOL | Electricity cost for air handlers used for cooling, in dollars, 2015 |
| 4 | DOLELAHUHEAT | Electricity cost for air handlers and boiler pumps used for heating, in dollars, 2015 |
| 5 | DOLELCOL | Electricity cost for air conditioning (central systems and individual units), in dollars, 2015 |
| 6 | DOLELLGT | Electricity cost for indoor and outdoor lighting, in dollars, 2015 |
| 7 | DRYUSE | Frequency of clothes dryer use |
| 8 | H2OHEATAPT | Water heater in apartment or other part of building |
| 9 | LGTINNUM | Number of light bulbs installed inside the home |
| 10 | LGTOUTNUM | Number of light bulbs installed outside the home |
| 11 | MONEYPY | Annual gross household income for the last year |
| 12 | NHSLDMEM | Number of household members |
| 13 | SOLAR | On-site electricity generation from solar |
| 14 | STORIES | Number of stories in a single-family home |
| 15 | SWIMPOOL | Swimming pool |
| 16 | TOTALBTU | Total usage, in thousand Btu, 2015 |
| 17 | TOTALBTUNEC | Total usage for other devices and purposes not elsewhere classified, in thousand Btu, 2015 |
| 18 | TOTALBTUSPH | Total usage for space heating, main and secondary, in thousand Btu, 2015 |
| 19 | TOTALBTUWTH | Total usage for water heating, main and secondary, in thousand Btu, 2015 |
| 20 | TOTALDOL | Total cost, in dollars, 2015 |
| 21 | TOTALDOLNEC | Total cost for other devices and purposes not elsewhere classified, in dollars, 2015 |
| 22 | TOTALDOLSPH | Total cost for space heating, main and secondary, in dollars, 2015 |
| 23 | TOTALDOLWTH | Total cost for water heating, main and secondary, in dollars, 2015 |
| 24 | TOTROOMS | Total number of rooms in the housing unit, excluding bathrooms |
| 25 | WASHLOAD | Frequency of clothes washer use |
| 26 | WINDOWS | Number of windows |

## References


1. U.S. Energy Information Administration, 2015 Residential Energy Consumption Survey. https://www.eia.gov/consumption/residential/reports.php
2. Bohlmann, J.A., Inglesi-Lotz, R.J.R.: Analysing the South African residential sector's energy profile. Renew. Sustain. Energy Rev. **96**, 240–252 (2018)
3. Hu, S., Yan, D., Guo, S., Cui, Y., Dong, B.: A survey on energy consumption and energy usage behavior of households and residential building in urban China. Energy Build. **148**, 366–378 (2017)
4. Dinesh, C., Makonin, S., Bajic, I.V.: Incorporating time-of-day usage patterns into non-intrusive load monitoring. In: Signal and Information Processing (GlobalSIP), 2017 IEEE Global Conference, pp. 1110–1114. IEEE (2017)
5. Gajowniczek, K., Ząbkowski, T.: Data mining techniques for detecting household characteristics based on smart meter data. Energies **8**(7), 7407–7427 (2015)
6. Perez-Chacon, R., Talavera-Llames, R.L., Martinez-Alvarez, F., Troncoso, A.: Finding electric energy consumption patterns in big time series data. Distributed Computing and Artificial Intelligence, 13th International Conference. AISC, vol. 474, pp. 231–238. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40162-1_25
7. Zhang, P., Wu, X., Wang, X., Bi, S.: Short-term load forecasting based on big data technologies. CSEE J. Power Energy Syst. **1**(3), 59–67 (2015)
8. Sheikhan, M., Mohammadi, N.: Neural-based electricity load forecasting using hybrid of GA and ACO for feature selection. Neural Comput. Appl. **21**(8), 1961–1970 (2012). https://doi.org/10.1007/s00521-011-0599-1
9. Zhang, M., Bai, C.: Exploring the influencing factors and decoupling state of residential energy consumption in Shandong. J. cleaner prod. **194**, 253–262 (2018)
10. Koprinska, I., Rana, M., Agelidis, V.G.: Correlation and instance based feature selection for electricity load forecasting. Knowl. -Based Syst. **82**, 29–40 (2015)
11. N. Amjady, F. J. E. C. Keynia, and Management, "Day-ahead price forecasting of electricity markets by a new feature selection algorithm and cascaded neural network technique," vol. 50, no. 12, pp. 2976–2982, 2009
12. Hu, Z., Bao, Y., Xiong, T., Chiong, R.: Hybrid filter–wrapper feature selection for short-term load forecasting. Eng. Appl. Artif. Intell. **40**, 17–27 (2015)
13. Jović, A., Brkić, K., Bogunović, N.: A review of feature selection methods with applications. In: 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015, pp. 1200–1205. IEEE (2015)
14. What's New in How We Use Energy at Home: Results from EIA's 2015 Residential Energy Consumption Survey (RECS). https://www.eia.gov/consumption/residential/reports.php
15. Residential Energy Consumption Survey (RECS) 2015 Household Characteristics Technical Documentation Summary. https://www.eia.gov/consumption/residential/reports.php
16. Han, J., Pei, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, Amsterdam (2011)