CLUSTERING HOUSEHOLD ELECTRICITY AND GAS LOAD PROFILES IN THE UK USING SMART METER AND SURVEY DATA

Zeynep Duygu Tekler[1]
[1]Department of Engineering Science, University of Oxford

ABSTRACT

The transition to Net Zero energy systems requires a comprehensive understanding of household energy consumption patterns. This understanding is crucial for developing simulation models that accurately reflect real-world building energy consumption and inform the design and implementation of effective energy policies and demand-side management interventions. This paper presents a real-world dataset collected from 139 UK households over a three-year period, consisting of smart meter and socio-demographic survey data. We apply machine learning techniques, including the K-means clustering algorithm, to identify and analyze distinct clusters for household gas and electricity load profiles. We also developed an XGBoost machine learning model to learn the relationship between household socio-demographics, building characteristics, and energy savings habits, using the model to interpret the clusters. Our findings reveal that the households' self-reported annual energy spend (i.e., utility bills) and the number of rooms (i.e., house size) are the two most important variables for inferring the households' electricity and gas clusters. Additionally, household annual income, specific housing types (i.e., Bungalow), and certain energy-saving habits also play smaller but significant roles during cluster assignment. The results offer a granular understanding of household energy consumption in the UK, facilitating the development of targeted energy policies and personalized demand response strategies. This research underscores the importance of integrating large-scale data acquisition with advanced analytics to support the implementation of equitable and effective Net Zero policies, ultimately contributing to smarter and more resilient urban energy systems.

Keywords: Electricity and Gas Load; Household Consumption; Large-scale Smart Meter Data; Machine Learning; Clustering

INTRODUCTION

Accurate modelling and simulation of household energy consumption have become helpful tools as many governments aim to meet mid-century Net Zero targets. In the UK, the residential sector alone accounts for over 30% of total energy consumption, with electricity making up 35-45% and gas 60-70% of this total (DESNZ, 2024). The availability of smart meter data has significantly enhanced these efforts, providing detailed insights into electricity and gas usage with unprecedented granularity. However, managing the vast volumes of smart meter data presents many challenges, including capturing the temporal and dynamic nature of energy consumption patterns.

Recent advancements in data analytics and machine learning have popularized clustering approaches for identifying representative load profiles that reflect diverse household energy use behaviors. Techniques such as Hierarchical Agglomerative Clustering (HAC), K-means, Fuzzy C-means (FCM), and Self-Organizing Map (SOM) are utilized to uncover intricate patterns in large datasets, creating more precise household clusters. These clustering algorithms defined clusters based on various factors, such as consumption habits and peak demand periods (Michalakopoulos, 2024). Accurate household load profiles obtained through these clustering approaches can help identify characteristics correlated with different energy usage patterns. This facilitates the design of tailored tariffs, and supports demand response initiatives, ultimately informing policy making.

Numerous studies have explored clustering load profiles using various models and metrics. For instance, McLoughlin (2015) presented a method for characterizing domestic electricity load profiles by comparing K-means, K-medoids, and SOM clustering approaches using the Davies-Bouldin Index (DBI). They also employed multinomial logistic regression to link profile classes with factors like dwelling type, occupants, and appliance characteristics. Similarly, Czetany (2021) analyzed smart meter data from Hungarian households and found that the K-means clustering approach was more effective than FCM and HAC for identifying daily and annual patterns, resulting in four distinct clusters. While smart meter data is commonly used for clustering load profiles with various algorithms, some studies have incorporated additional sociodemographic surveys and activity data. Tang (2022) identified twelve clusters for weekends and weekdays, mapping the relationship between load patterns and key socioeconomic features like age and education level. Sartre-Meloy (2020) conducted an in-depth analysis combining household time-use activity data to further examine these profiles during peak evening usage periods on weekdays.

Despite significant efforts in household load clustering, identifying representative clusters remains challenging. Clustering results are highly sensitive to various methodological factors, especially when analysing load profiles. Additionally, while clustering can be used to perform data segmentation, the interpretability of these clusters can be difficult. For applications such as demand response, tariff design, and understanding energy usage behaviors, it is crucial that clusters are explainable through descriptive data about buildings and their occupants. Advanced methods or models may be necessary to elucidate these clusters more effectively. Moreover, although many clustering analysis have been conducted for electricity profiles within households, gas load profiles have received comparatively less attention, particularly when analysed with electricity data.

This paper addresses these challenges by presenting a clustering methodology for household electricity and gas load profiles, utilizing longitudinal data from smart meters and household sociodemographic information. Additionally, the developed clusters and the features contributing to these diverse clusters were explained through a machine learning model trained on household socio-demographics, building characteristics, and energy-saving habits. The results provide valuable insights into different household energy usage patterns, enabling the design of customized demand response strategies. These findings offer valuable insights for policymakers, aiding in the development of effective energy management strategies aimed at achieving Net Zero targets.

METHOD

The Energy Demand Observatory and Laboratory

The Energy Demand Observatory and Laboratory (EDOL) represents a significant UK energy data infrastructure investment. This initiative aims to provide a longitudinal, disaggregated, and flexible resource of UK residential energy data, available to scientists, industry, and policymakers. EDOL is set to develop innovative and cost-effective smart data solutions for large-scale energy data collection. The EDOL Observatory will involve 2,000 representative UK households, and Laboratories will create an environment for testing interventions, and additional instruments. Different technology trials, and engagement strategies will also be evaluated in these Laboratories, with the Observatory as a control group.

The Sample and Data Collection

The sample is part of an EDOL pilot lab consisting of 139 households equipped with smart meters and recruited from an online market research panel. Participants are chosen to be demographically and geographically representative of Great Britain. Recruitment quotas by gender and region were also applied to enhance the sample's representativeness of the GB population.

The smart meter data used in this study includes half-hourly electricity usage from 139 households and gas consumption from 111 of these households over a three-year period (December 2020 to April 2024 for electricity and February 2021 to April 2024 for gas). This data, extended from METER dataset (Grunewald, 2020), is collected via mobile signals and accessed through the Data Communications Company (DCC) and Hildebrand Ltd. Glow service. To access household smart meter data, participants provide the 16-digit alphanumerical GUID from their in-home display, which is validated against their postcode to uniquely identify their smart meter. Once validated, participants received a follow-up email inviting them to join the study by agreeing to terms and conditions for accessing and processing their smart meter data.

In addition to the smart meter data, an occupant from each household is instructed to complete an online survey consisting of 26 questions covering household socio-demographics, building characteristics, as well as energy usage and saving habits. Table 1 presents a selected list of questions from the survey.

Data Processing

To clean the raw electricity and gas consumption data and extract relevant features for clustering, a series of data processing steps are performed.

The first step removes any duplicate records. Next, any negative gas consumption readings were nullified and imputed with the last recorded valid readings using a forward filling approach. This step was not applied to the electricity consumption readings, as negative values were possible due to the presence of solar photovoltaic panels installed in some of the participating households. The third step entailed identifying any abnormal consumption readings from the electricity and gas consumption data and imputing them with a forward filling approach. A reading was defined as abnormal if it was both unusually large and occurred infrequently within a short time window. Less than 1% of the data is affected by these steps. The fourth step removes the consumption data captured during the weekends, as weekend consumption patterns can be materially different and negatively impact the clustering performance.

Based on the remaining data, the following processing steps were performed separately for each household to extract relevant features for load clustering. The household consumption data was first aggregated from 30-minute to 1-hour intervals. Then, the household's hourly mean consumption was calculated, resulting in 24 features (one feature for each hour). These features provide information about the shape of the households' consumption patterns over time. Following the feature extraction approach proposed by Michalakopoulos (2024), two additional processing steps are performed to extract more informative features for identifying distinct load profiles. Firstly, the peak consumption readings recorded during six specific periods of the day (peak early morning: 5-10 AM, peak morning: 10 AM-2 PM, peak noon: 2-5 PM, peak evening: 5-9 PM, peak night: 9 PM-12 AM, peak late night: 12-5 AM) were extracted to obtain six additional features. These features provide information about the households' peak consumption while accounting for their temporal differences. Next, several descriptive statistics (mean, standard deviation, minimum, maximum, 25th percentile, median, 75th percentile) were also calculated from the household's hourly average consumption, resulting in a total of 37 features (24 + 6 + 7) for load clustering. The inclusion of these statistical measures provides a comprehensive summary of the households' consumption data, capturing both central tendencies and variability. Lastly, to ensure that the magnitude of each extracted feature followed the same scale and is weighted equally during the clustering process, a min-max scaling step was separately applied to each feature across all households, resulting in these features taking values between 0 and 1.

Data Clustering Approach

The clustering approach adopted in this study is the K-means clustering algorithm due to its popular use in similar studies (Michalakopoulos, 2024) and its implementation simplicity (Tekler, 2020). The algorithm begins by randomly assigning K data points as the initial centroids, where K is user-defined and indicates the number of clusters to be found within the data. Next, by calculating the sum of squared distances between each data point and the centroids, each data point is assigned to a particular cluster based on its nearest centroid. Following this, a new set of cluster centroids is calculated based on the average distance of the data points assigned to their respective clusters, before repeating this process until the cluster assignments converge.

Since it is infeasible to know in advance the number of clusters within the electricity and gas consumption data, we identified the optimal number of clusters by evaluating a range of K values using a combination of popular clustering validity indicators (CVIs) and identifying the cluster number that resulted in the best clustering performance across all CVIs. The CVIs used in this study are the Calinski-Harabasz Index, Davies-Bouldin Index, and Silhouette Score, each of which involves different methods of calculating the compactness within each cluster and the separation between different clusters.

Cluster Interpretation using XGBoost

After clustering the households' electricity and gas consumption data based on the optimal number of clusters identified, it is useful to provide an interpretation for these clusters and identify the important factors contributing to the households' cluster assignments. This is achieved by first merging the households' cluster assignments with their survey data, which contains information about households' socio-demographic information, building characteristics, as well as energy usage and saving habits.

Following which, an eXtreme Gradient Boosting (XGBoost) classifier is fitted to the households' survey data to infer the cluster assignments. The XGBoost model is chosen due to its superior predictive performance and computational efficiency. The training process involves fitting the XGBoost classifier on 80% of the data and evaluating the model on the remaining 20% to ensure the model generalizes well to new households.

Macro-average accuracy and macro-average F1 score are used to evaluate the XGBoost classifier's performance, rather than their micro-average versions, to ensure the model correctly assigns households uniformly across all clusters. Once a model with good classification performance is obtained, the most important features for cluster assignment are identified by calculating the SHapley Additive exPlanations (SHAP) values for each feature. These values are computed by evaluating the contribution of a particular feature to the model's predictions, determined by the difference in prediction when the feature is included versus excluded, averaged across all feature subsets.

RESULTS AND DISCUSSION

Clustering Results

As described in the previous section, the optimal number of clusters for both the electricity and gas consumption data is determined by evaluating the clusters obtained across a range of cluster numbers using the three CVIs discussed in the Methods section. Based on Figure 1, it can be observed that the three CVIs are in complete agreement with each other regarding the optimal number of clusters for both the electricity (2 clusters) and gas (2 clusters) load profiles.

By dividing the households' electricity load profiles into two distinct clusters, it was found that the first cluster comprises 123 households (88.5%), while the second cluster comprises 16 households (11.5%). By visualizing the households' aggregated electricity consumption data based on their assigned clusters and calculating their respective cluster means, we can observe from Figure 2 that the households in Cluster 1, on average, follow a unimodal distribution, while those in Cluster 2 follow a multimodal distribution.

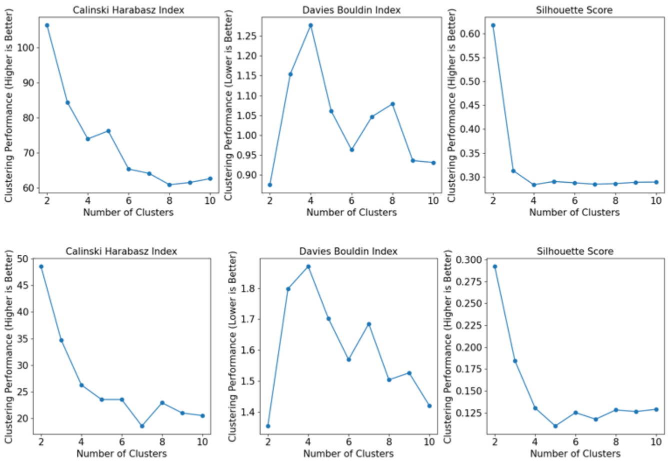

Figure 1: *Clustering performance for electricity consumption data (top) and gas consumption data (bottom). The quality of the clusters is iteratively evaluated over a range of clusters (2 to 10) using the Calinski and Harabasz index (left), Davies-Bouldin index (middle), and Silhouette score (right).*

More specifically, the average household electricity consumption in Cluster 1 starts with a U-shaped curve, gradually falling from midnight and reaching its minimum value at 5 AM before rising again between 7 to 8 AM as the occupants wake up and begin their day. From 8 AM onwards, the average electricity consumption remains fairly constant before rising significantly from 4 PM and peaking at 6 PM. This significant rise in electricity consumption might be attributed to the increased use of household appliances as occupants return home, seek entertainment, or begin preparing dinner. After 6 PM, the electricity consumption for Cluster 1 gradually decreases as the occupants slowly turn in for the night.

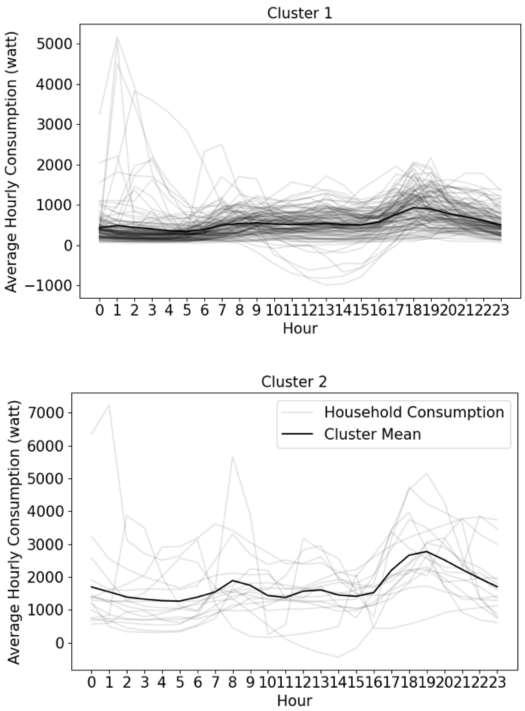

Figure 2: *Household aggregated electricity consumption data for Cluster 1 (top) and Cluster 2 (bottom), together with their cluster means.*

Cluster 2 follows a similar pattern to Cluster 1 with two key differences. The first difference is the presence of three consumption peaks occurring at 8 AM, 12:30 PM, and 7 PM, which coincide with meal times. The second difference is the significantly higher average consumption compared to Cluster 1.

Similarly, the two gas consumption clusters contain 61 household (55.0%), and 50 households (45.0%) respectively. By visualizing the households' gas load profiles based on their assigned clusters and calculating their respective cluster means, we can observe from Figure 3 that the households in both clusters follow a bimodal distribution.

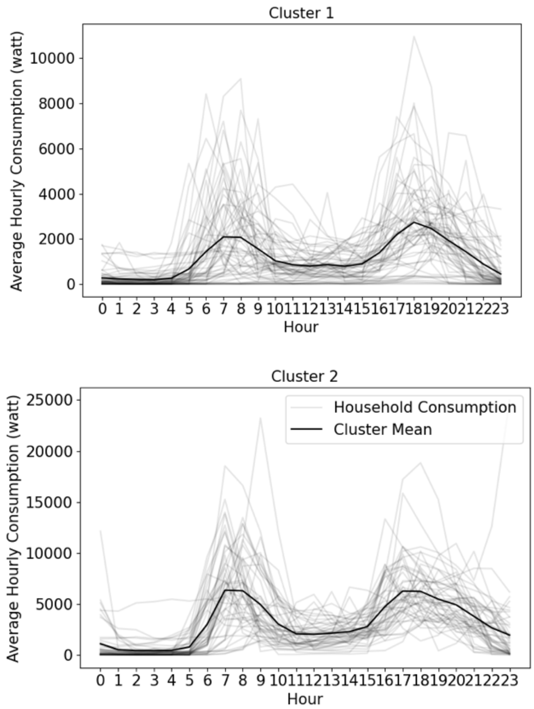

Figure 3: *Household aggregated gas consumption data for Cluster 1 (top) and Cluster 2 (bottom), together with their cluster means.*

More specifically, the average household gas consumption values in both clusters are at their minimum levels between midnight and 4 AM before rising significantly and peaking between 7 and 8 AM as the occupants wake up and begin their day. After 8 AM, the gas consumption profiles for both clusters follow a U-shaped curve, falling to a trough between 11 AM and 3 PM before rising sharply again to the second peak of the day at 6 PM. After 6 PM, the household gas consumption for both profiles gradually decreases as the occupants turn in for the night.

The two consumption peaks observed in both clusters might be attributed to a range of activities by the occupants, such as cooking, dishwashing, and cleaning, all of which consume gas either directly or indirectly. It should also be noted that while Cluster 2 follows a very similar pattern to Cluster 1, the key difference between the two clusters is in the peak gas consumption, with Cluster 2 having a much higher peak gas consumption (>5,000 watts) compared to Cluster 1 (~2,000 watts).

Cluster Interpretation using surveys via XGBoost

After obtaining the cluster assignments for each household, these cluster assignments are interpreted by fitting an XGBoost classifier on the survey data containing information about the households' socio-demographic information, building characteristics, energy usage and saving habits, as shown in Table 1.

Due to the significant imbalance between the number of households assigned to Cluster 1 (88.5%) compared to Cluster 2 (11.5%) for the electricity load profiles, the XGBoost classifier trained on this dataset will be skewed towards the majority class (i.e., Cluster 1), leading to misinterpretations regarding the cluster assignments for the minority class (i.e., Cluster 2). To resolve this issue, an additional random oversampling step is included, which randomly samples the minority class with replacement to rebalance the number of households between the two clusters before training the XGBoost classifier on the oversampled household survey data. While more advanced oversampling techniques exist (e.g., Synthetic Minority Over-sampling Technique and Adaptive Synthetic Sampling), an empirical evaluation of these methods showed a drop in the model's performance, aligning with findings from past empirical studies (Low, 2020). By including the oversampling step, the resulting classification model achieved an excellent test performance with a macro-average accuracy of 0.960 and an F1 score of 0.854 for the electricity cluster assignments. Additionally, the classification model for gas consumption reported a macro-average accuracy score of 0.784 and a macro-average F1 score of 0.783 on the test data.

Electricity Profiles, Cluster 1 (n=123) and Cluster 2 (n=16)

Using the classification model trained on the survey data with electricity cluster assignments, the top 10 most important features contributing to cluster assignment for electricity consumption are depicted in Figure 4. According to Figure 4, the most important feature is the household's self-reported energy spend, where a lower energy spend (represented by blue dots on the SHAP summary plot) results in a negative SHAP value, increasing the cluster propensity towards Cluster 1. Analysing the descriptive statistics in Table 1, households with lower reported energy spend tend to be classified into Cluster 1. Specifically, 8% of Cluster 1 participants report to be spending less than £500 annually on energy, while Cluster 2 shows no participants within this lowest spend bracket. Furthermore, 78% of Cluster 1 participants reportedly spend between £500 and £1499, compared to only 19% in Cluster 2.

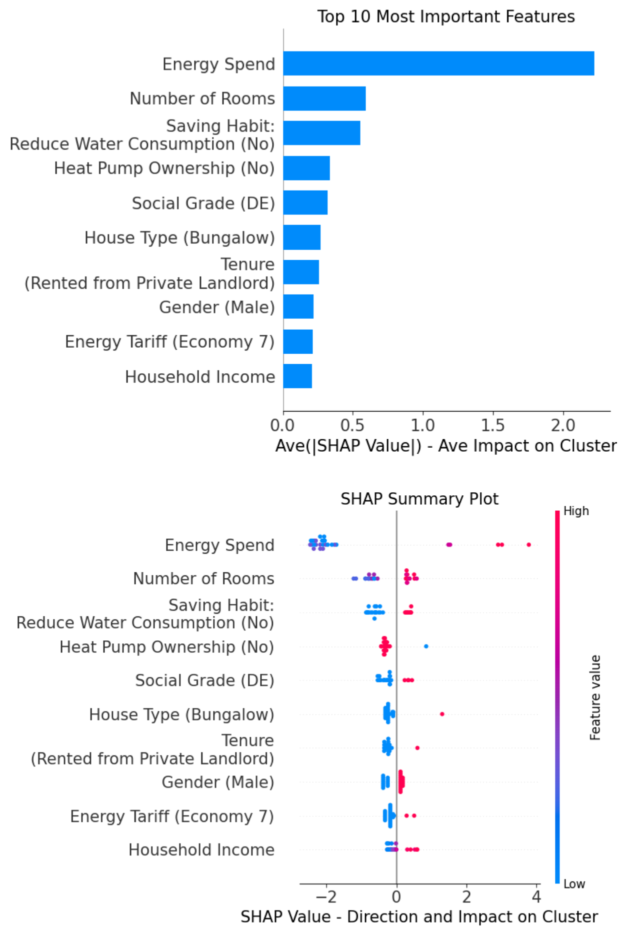

Figure 4: *The feature importance plot (top) and SHAP summary plot (bottom) for the top 10 most important features for household electricity cluster assignment.*

The second most important feature is related to the number of rooms in the household, where fewer rooms correlate with a higher likelihood of being assigned to Cluster 1. For Cluster 1, a moderate portion (22%) of households have 6+ rooms, whereas Cluster 2 has a higher percentage (50%) for the same category. This indicates that households with fewer rooms, which are often indicative of smaller home size, tend to be assigned to Cluster 1.

Variables	Categories	Electricity Sample (%) (n=139)	E_Cluster 1 (%) (n=123)	E_Cluster 2 (%) (n=16)	Gas Sample (%) (n=111)	G_Cluster 1 (%) (n=61)	G_Cluster 2 (%) (n=50)
Gender	Male	59	59	63	54	44	66
	Female	41	41	38	46	56	34
Age	18 to 44	13	12	19	13	16	8
	45 to 54	22	21	31	23	28	16
	55 to 64	26	26	25	24	26	22
	65 or older	42	41	25	40	30	54
Num of Adults	Up to 2	88	89	81	89	86	90
	3+	12	11	19	11	11	10
Work Status	Not working	14	13	25	13	15	12
	Retired	38	41	19	41	30	54
	Working full-time	36	36	38	34	41	26
	Working part-time	12	11	19	12	15	8
Social Grade	AB	41	41	33	41	36	46
	C1	27	29	18	26	26	26
	C2	13	11	25	13	11	14
	DE	17	17	19	20	26	12
Household Income	>£28000 or unknown	42	44	19	41	49	30
	£28001 to £48000	33	32	37	37	34	38
	£48001 +	25	24	44	22	17	32
Own Heat Pump	No	96	100	69	100	100	100
	Yes	4	0	31	0	0	0
Tenure	Own	83	84	81	86	77	86
	Rent	17	16	19	14	23	14
House Type	Bungalow	7	6	19	6	10	6
	Flat/apartment	14	16	0	9	16	0
	House, detached	27	25	44	25	16	32
	House, semi-detached	27	28	19	33	28	40
	House, terraced	24	24	19	26	30	20
Num of Rooms	>4	51	55	19	47	61	32
	5	24	24	31	25	23	28
	6+	25	22	50	27	16	40
Energy Spend	>£500	7	8	0	6	11	0
	£500-£1499	72	78	19	70	76	68
	£1500 +	23	14	82	22	13	32
Habit: Reduce Water Consumption	No	37	35	56	34	33	36
	Yes	63	65	44	66	67	64
Habit: Set Thermostat Low	No	46	45	56	41	31	52
	Yes	54	55	44	59	69	48
Habit: Wash at Cold Temperature	No	58	59	50	55	48	64
	Yes	42	41	50	45	52	36

Table 1. *Selected descriptive statistics for each survey variable and its respective categories across the electricity and gas household samples, with two clusters for each. Some categories have been merged to avoid data disclosure.*

Lastly, households that adopt habits of reducing their water consumption are also more likely to be assigned to Cluster 1. These results align with the load profile analysis conducted in the previous subsection, where households in Cluster 1, which have lower average energy consumption compared to households in Cluster 2, tend to spend less on energy, have lesser number of rooms (proxy for house size), and adopt energy conservation habits like reducing water consumption.

Gas Profiles, Cluster 1 (n=61) and Cluster 2 (n=50)

Using the classification model trained on the survey data with gas cluster assignments, the top 10 most important features contributing to cluster assignment for gas consumption are depicted in Figure 5. Based on Figure 5, the top two most important feature for gas cluster assignment are the household's annual self-reported energy spend and the number of rooms. More specifically, a higher energy spend and more rooms (represented by red dots on the SHAP summary plot) result in positive SHAP values, which increases the households' cluster propensity towards Cluster 2. From Table 1, it can be observed that households in Cluster 2 have a higher reported energy spend, with 68% spending £500-£1499 and 32% spending more than £1500 compared to Cluster 1 where 76% spend £500-£1499 and only 13% spend more than £1500. Additionally, households in Cluster 2 typically have more rooms, with 40% having 6+ rooms compared to 16% in Cluster 1.

It should be noted that the insights from this subsection do not imply a causal relationship between these factors and the households' energy consumption. Rather, this analysis is meant to identify the important factors that can be used to infer the household's electricity and gas consumption profiles. A secondary and more practical benefit of developing such a model for inferring the households' load profile is that it can help supplement existing data collection efforts where installing smart meters may be technically challenging. By being able to use surveys to infer households' load profiles, these existing data gaps can be addressed until more advanced data collection techniques or sensors are developed.

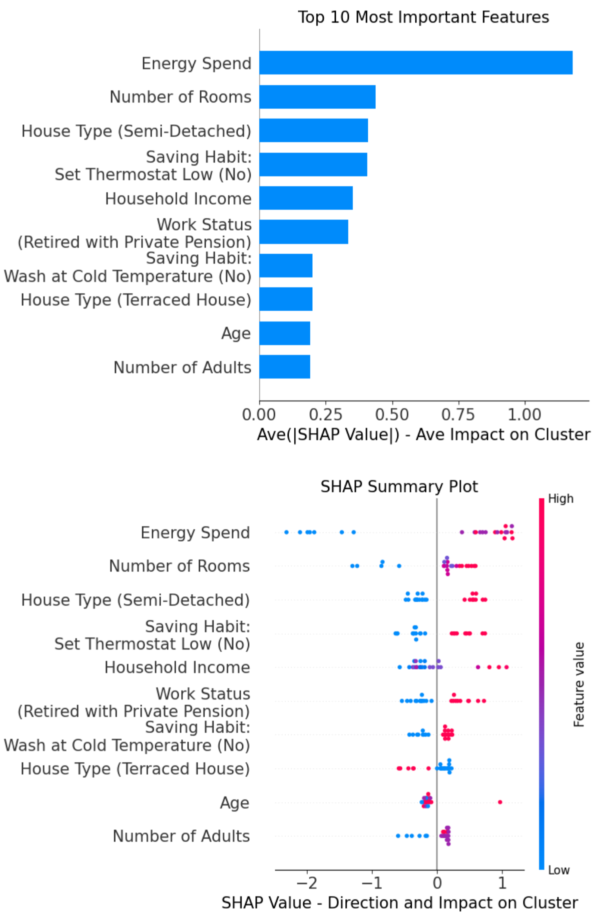

Figure 5: *The feature importance plot (top) and SHAP summary plot (bottom) for the top 10 most important features for household gas cluster assignment.*

CONCLUSIONS

This study analysed electricity and gas consumption profiles of UK households using over three years of smart meter data and household surveys. Employing the K-means algorithm, two distinct clusters for both electricity and gas were identified using three CVIs. Subsequently, we employed an XGBoost machine learning model to interpret these clusters, identifying key features that affect cluster assignments. The findings reveal that annual energy spend and the number of rooms are the most significant factors, with household income, house types, and energy-saving habits also playing notable roles. These insights can facilitate in designing demand response strategies, support energy modelling, and contribute to NetZero goals. They are also valuable for developing new household energy archetypes that reflect distinct consumption patterns, informing tailored energy efficiency programs, localised microgrids, and long-term energy forecasts for sustainable urban planning and grid management. Future work will expand the analysis to a larger sample size, with ongoing data collection at EDOL offering opportunities to support demand response analysis.

ACKNOWLEDGEMENTS

This research was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant EP/X00967X/1.

REFERENCES

UK DESNZ (2024). Subnational Electricity and Gas Statistics 2022. Accessed: 10 July 2024.

Michalakopoulos, V., Sarmas, E., Papias, I., Skaloumpakas, P., Marinakis, V., & Doukas, H. (2024). A machine learning-based framework for clustering residential electricity load profiles to enhance demand response programs. *Applied Energy*, *361*, 122943.

McLoughlin, F., Duffy, A., & Conlon, M. (2015). A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied energy*, *141*, 190-199.

Czétány, L., Vámos, V., Horváth, M., Szalay, Z., Mota-Babiloni, A., Deme-Bélafi, Z., & Csoknyai, T. (2021). Development of electricity consumption profiles of residential buildings based on smart meter data clustering. *Energy and Buildings*, *252*, 111376.

Tang, W., Wang, H., Lee, X. L., & Yang, H. T. (2022). Machine learning approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data. *Energy*, *240*, 122500.

Satre-Meloy, A., Diakonova, M., & Grünewald, P. (2020). Cluster analysis and prediction of residential peak demand profiles using occupant activity data. *Applied Energy*, *260*, 114246.

Grünewald, P., & Diakonova, M. (2020). METER: UK Household Electricity and Activity Survey, 2016–2019. *UK Data Service: Colchester, UK*.

Tekler, Z. D., Low, R., Gunay, B., Andersen, R. K., & Blessing, L. (2020). A scalable Bluetooth Low Energy approach to identify occupancy patterns and profiles in office spaces. *Building and Environment*, *171*, 106681.

Low, R., Cheah, L., & You, L. (2020). Commercial vehicle activity prediction with imbalanced class distribution using a hybrid sampling and gradient boosting approach. *IEEE Transactions on Intelligent Transportation Systems*, *22*(3), 1401-1410.