Creating and Characterising Electricity Load Profiles of Residential Buildings

James Fitzpatrick[1], Paula Carroll[1], and Deepak Ajwani[2]

[1] Quinn School of Business, University College Dublin
{james.fitzpatrick1@ucdconnect.ie, paula.carroll@ucd.ie }
[2] School of Computer Science, University College Dublin
deepak.ajwani@ucd.ie

Abstract. Intelligent planning, control and forecasting of electricity usage has become a vitally important element of the modern conception of the energy grid. Electricity smart-meters permit the sequential measurement of electricity usage at an aggregate level within a dwelling at regular time intervals. Electricity distributors or suppliers are interested in making general decisions that apply to large groups of customers, making it necessary to determine an appropriate electricity usage behaviour-based clustering of these data to determine appropriate aggregate load profiles. We perform a clustering of time series data associated with 3670 residential smart meters from an Irish customer behaviour trial and attempt to establish the relationship between the characteristics of each cluster based upon responses provided in an accompanying survey. Our analysis provides interesting insights into general electricity usage behaviours of residential consumers and the salient characteristics that affect those behaviours. Our characterisation of the usage profiles at a fine-granularity level and the resultant insights have the potential to improve the decisions made by distribution and supply companies, policy makers and other stakeholders, allowing them, for example, to optimise pricing, electricity usage, network investment strategies and to plan policies to best affect social behavior.

Keywords: Smart-meter · Load-profiling · Time series clustering.

1 Introduction

Accurately characterizing the daily load profile of electricity usage has the potential to considerably improve the decision making for electricity suppliers and distributors, customers, policy makers and various other stakeholders. For instance, it can help suppliers to optimise pricing, distributors to develop better distribution strategies, manage the peak demand and find ways to flatten the peak and it can support policy makers to align climate action plans with cleaner energy initiatives.

In particular, a careful analysis of the smart meter time series data, with a view to learn insights for characterizing the daily load profile of residential customers has potential to assist the various stakeholders in taking a data-driven

approach to their decision making. However, extracting these insights and understanding the connections between electricity usage, the dwelling and the consumer behaviour is non-trivial. The smart meter time series data at the individual dwelling level are noisy but when aggregated to groups of users evaluated over time can reveal patterns of behaviours. Such patterns, or representative load profiles, indicate when the peak demand may occur for groups of customers, and are used by electricity market operators to schedule generation to meet demand. There are opportunities to encourage users to moderate their electricity usage patterns so as to reduce aggregate peak demand, but first we need to develop an understanding of the representative load profiles.

In this paper, we consider the case when the user remains in control of their electricity usage, rather than an intelligent energy management system. We take the perspective of an electricity supplier or policy maker wishing to understand residential consumers electricity usage. We base our work on a smart meter customer behaviour trial which was carried between 2009 and 2011 [6]. Participants retained total autonomy over the scheduling of their electricity usage during the trial. For each participant, a survey was carried out before and after the installation to determine the characteristics of the building construction and the household composition, as well as their attitudes to the electricity usage and expected benefits of a smart meter. This multivariate data-set of smart meter time series and survey responses provides a unique opportunity to study the relationship between the characteristics of a dwelling and its electricity consumption pattern, when the consumption information is accessible to the users. Policy makers would be interested know which of the survey features best explain consumer electricity usage patterns. Analysis of survey responses using explainable techniques provides actionable insights that can be targeted in electricity efficiency programmes.

The smart meter usage data are stochastic and high-dimensional. In order for actors in the electricity market to incorporate these data into data-driven decision-making processes, we must consider how to reduce the dimensionality, model the data and extract useful insights. In this paper, we explore appropriate schemes for carrying this out and to answer the following research questions:

1. Can we create representative load profiles for clusters of smart meter users based on time series electricity usage?
2. Can we characterise the cluster representative load profile using information in the survey?
3. What insights do the cluster characteristics provide to support the development of climate action and electricity usage incentives?

We address these research questions by first performing a careful clustering of the normalized electricity load time-series to learn the behavioural daily patterns of residential customers. Then, we learn a classification model to map the survey features to the clusters. In the process, we focus on the importance of various survey features and learn crucial insights from this analysis. Our insights can be valuable for distribution and supply companies, policy makers and various

stakeholders in the energy business. For instance, we learn that one of the most important features in predicting the daily load cluster is "how strongly they feel that they can convince other occupants of the building to reduce their energy usage." Given that the survey has many detailed characteristics of the building and the household, the consistent importance of this feature across many different classification models is surprising. This, itself, is an important finding for a country like Ireland, which has traditionally struggled to get value out of retro-fitting houses for energy efficiency improvements [16]. Our finding suggests that a marketing campaign to change the attitudes of people towards energy efficiency may be effective in modifying the daily usage pattern of residential customers.

Outline. This rest of the paper is structured as follows: Section 2 describes the related work, Section 3 details the structure of the time series and describes the survey data and the problem outline, Section 4 concerns the clustering of the time series and the creation of aggregate load profiles and the process of mapping survey responses to their corresponding time series cluster, Section 5 illustrates the experimental results and provides an exposition on these results and Section 6 presents our conclusions.

2 Related Work

In this section, we review the literature related to the usage of time series clustering for smart meter data. We briefly survey (i) the techniques developed for time-series clustering in general, then (ii) cover the work related to the usage of time-series clustering for smart meter electricity data with a specific focus on the Irish customer behaviour trial data and (iii) characterization of smart meter load profile of residential users based on the attributes of the residential building.

Time-series clustering. Clustering of time-series data has been an active area of research over the last few decades and many good techniques have been developed (c.f. [2, 11, 17] for surveys and [13] for some recent work). The challenge in clustering the smart meter data stems from:

1. Electricity usage time series is inherently noisy. Such noise emerges naturally from the stochasticity of human lifestyles, but also from climactic and weather conditions, and even possibly the purposeful injection of noise to ensure privacy [9]).
2. Time-series differ in length. While this challenge is typically addressed using dynamic time warping measures (as highlighted in the review [7]), these methods are sensitive to noise, making the resolution of the first challenge even more challenging.
3. We are not interested in clustering based on the total usage, but in identifying different shapes of the standardised time-series, corresponding to the different daily patterns of the consumers. The tasks of clustering based on total

usage and learning to forecast the load based on attributes of the household are relatively easier, our task of learning the daily pattern of a household is significantly more difficult.
4. For the public policy bodies and industry analysts to be able to act on the models and the resultant insights, the clustering of the time series and the mapping of the survey data to the clusters should be as interpretable as possible.

Analysis of smart metering infrastructure. The installation of smart metering infrastructure in recent years has sparked interest in the desire to develop methods to draw insights from the data that is being collected. This includes not only electrical smart meters, but also water and gas smart meters [4, 5, 12, 15]. Understanding how groups of consumers behave makes it possible to plan infrastructure projects, develop pricing strategies and identify anomalous behaviours. Naturally, clustering can be performed trivially for cases of separating commercial and industrial consumers from residential consumers, as well as by grouping by consumption magnitude. In contrast to most existing works, our focus is on the considerably more difficult task of learning the behaviour-based clusters to better understand how consumers consume. Such a clustering reveals the different daily usage patterns of residential customers and enables us to learn which features of the buildings, households and people's attitudes best discriminate between the different clusters, revealing crucial insights for policy makers.

Clustering the time-series daily usage pattern from household. While there is considerable body of work on clustering residential electriciy customers using load time series (see e.g., [14]), there is very little work on correlating it with the features of the household and building, leave aside our goal of inferring the usage pattern from the household and building features. Lavin and Klabjan [10] constructed mean normalised daily energy profiles for each meter in their data-set of commercial and industrial buildings in the United States. They noted that the daily usage pattern could be used to determine the work schedule in the commercial buildings. Note that our focus is on the significantly more challenging task of learning the behavioural usage pattern from the household and building features. Alonso et al. [1] focused on scalable clustering of the time-series by reducing their time series representation to autocorrelation coefficients. They showed that the clusters that they obtained correlated well with the geo-demographic data related to the class and social status of individuals. In contrast, we take the study to the next level and attempt to infer the usage pattern from a range of features and identify the features that are most discriminatory. In Flath et al. [8], standard normalised daily load and weekly profiles for nine scenarios recognised by the German energy industry were computed as features from time series data. These previously known load profiles were used to perform clustering of the time-series data from a pricing perspective. However, they do not seek to explain the underlying characteristics of the buildings to which the smart meters are connected. Also, in contrast to their work, we identify the importance of each feature in identifying the usage patterns without any assumptions a priori.

Analysis of Irish customer behaviour trial data. There has also been some work on the analysis of the Irish customer behaviour trial data [6] that we use in this study. Carroll et al. [5] derived statistical features from the time series over a period of six months and attempted to solve the problem of inferring composition of a household living in a building based on the features that characterise the electricity usage behaviour of the smart meter time series. In contrast, this paper focuses on the significantly more challenging task of learning the usage behaviour from the features obtained using the associated survey.

A closely related work is that of McLoughlin et al. [12], who performed subsequence clustering of the CER [6] residential electricity smart meter time series by considering the first six months of recordings for each meter using self organising maps. However, they focused on the regression models and more crucially, ignored the features corresponding to how often the household appliances were used (only using if appliances such as washing machine were present in the household) and the attitudes of the occupants towards energy saving and metering measures. In contrast, we found that these features were the most important in discriminating between the different usage patterns of household customers. Azaza and Frederik [3] analyse the same data-set, using self-organizing maps and hierarchical methods, clustering the time series using daily mean energy usage profiles. But they only attempt to understand each cluster from an energy usage perspective, not a building composition perspective. In contrast, our study addresses the challenging task of learning the clusters of daily usage patterns from the accompanying survey data.

3 Smart Meter Characterisation and Classification Problem

In this work we are concerned with the creation of electricity load profiles for residential electricity consumers. Associating a load profile to each customer allows distributors and suppliers to anticipate expected user behaviour, plan infrastructure and targeted interaction strategies accordingly.

We first perform a clustering of the residential consumers into relatively large and roughly equal-sized clusters based on a transformation of their smart meter time series electricity usage. We then construct a mapping from the survey responses to these clusters to characterise the clusters. Finally we analyse the load profiles for these clusters and the salient survey questions to better understand the cluster behaviour and potential for targeted electricity savings interventions.

Dataset For this study, we use a data-set [6] obtained from a customer behaviour trial that was carried out between 2009 and 2011. This trial was carried out in a range of Irish residential and commercial buildings to observe the response to the installation of smart meters. Participants retained total autonomy over the scheduling of their electricity usage during the trial. For each participant, a survey was carried out before and after the installation to determine the

characteristics of the building construction, the composition of the household, as well as attitudes to the energy usage and expected benefits of a smart meter.

The trial includes 6445 participants, of which 4225 were residential participants. From these residential participants, we filtered out the ones with suspected instrumentation faults as well as those for whom incomplete survey responses could not be reasonably imputed. This resulted in a total of 3670 participants that were considered for our work. Each residential smart meter is assigned to one building, representing a single household.

The smart meter time series data was collected at a half-hour granularity, that is, the power consumed over each half hour interval for the duration of the study was recorded for each participant. This corresponds to 48 time slots per day, over the course of 535 days, a univariate uniformly-sampled sequence. Some time series, however, were incomplete, meaning that they are not all of the same length; they did not begin or terminate at the same time as those that extended over the entire duration. For each participant i, therefore, we have a real-valued vector $X_i \in \mathbb{R}^{d_i}$. The vast majority of these univariate time series have more than ten thousand elements.

For each participant i, there is a unique smart meter time series X_i as well as a unique survey response Z_i, forming a complete data-set $\mathcal{D} = \{(X_i, Z_i)\}_{i=1}^{3670}$. Each residential participant completed a survey prior to and subsequent to the eighteen month trial. For our analysis, we only retain responses from the pre-survey questionnaire and only if they concern the household composition (the number of people who live in the household), the characteristics of the building or its contents, or if they indicate the attitude of the respondent to the expected outcome of the trial. Questions that have categorical answers are one-hot encoded and questions that admit ordinal responses are normalised by the maximum possible value, or recorded value, if there is no maximum. This results in a 110-dimensional response vector $\hat{Z}_i \in R^{110}$ to be associated with each smart meter time series.

4 Methodology

Extracting clusters from the data is equivalent to finding a label y_i for each of the pairs (X_i, Z_i). In this section we outline the feature extraction methods we use to find a fixed length feature vector $\hat{X}_i$ to characterise each smart meter time series and cluster them into pairs $(\hat{X}_i, y_i)$. We then discuss how, having constructed feature vectors $\hat{Z}_i$ from the survey data, we find some model $p(y_i|\hat{Z}_i; \theta)$.

4.1 Time Series Clustering

In order to derive insights from the smart meter time series upon which decisions can be made, they must be reduced significantly in dimension. It follows that it is desirable to construct a small number of clusters for which analysis can be carried out. This amounts to using unsupervised methods to determine some mapping $f : \hat{X} \to k$, where $k \in \{\{0, 1, 2\}, \{0, 1, 2, 3\}, \{0, 1, 2, 3, 4\}\}$ and $\hat{X} \equiv \{\hat{X}_i\}_{i=1}^{3670}$.

Three major paradigms are recognised for the clustering of time series data: whole-series (raw) clustering, extracted feature clustering and model-based clustering [11]. These residential electricity usage time series, driven by stochastic variables such as local weather conditions and human activities, are subject to a significant degree of noise, making the first of these approaches undesirable for clustering. In addition, it is preferable that the clustering of the time series is easily interpretable, so that decisions made on the basis of the generated clusters are reliable, enabling the public policy bodies and analysts to act on the resultant models. It is, therefore, desirable to compute a feature representation that captures the behaviour of each time series and its peculiarities.

For each time series X_i we know the mapping $g_i : X_i \to \{0, 1, ..., 47\}^{m_i}$, where m_i is the number of days for which observations of the meter i were made. That is, we have an exact mapping between each recorded power consumption value and the time of day at which it was recorded. We also know the correspondence between each measurement and the day and year it was recorded. This allows us to construct fixed-length, representations of the load corresponding to fixed time periods. Consider, for example, that a smart meter is observed n times per day at regular intervals over a period of m days, then we can represent each measurement in a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$. Such a representation contains exactly the same information as the one-dimensional representation, but we can reduce it to obtain the mean energy usage per time slot according to:

$$\mathbb{R}^n \ni \hat{X}_j = \frac{1}{m} \sum_{i=1}^{m} X_{ij}.$$

We can also construct similar features, in order to take into account the differences in behaviour that can be observed during weekdays and weekends, or on a weekly/monthly basis. These representations are static and can be easily used as feature vectors for static clustering algorithms.

In this work, clustering was performed using the k-means clustering algorithm; various clustering algorithms were tested, such as agglomerative and other density-based methods, but k-means produced the most well-separated clusters, as indicated by computation of Silhouette indices. A variety of static representations of the time series data, such as those discussed in the previous paragraph, were chosen as the feature vectors upon which the clustering was performed. We proceeded with an ℓ_2 norm as a dissimilarity measure. In order to determine the number of clusters, trial clusterings were performed for three, four and five clusters, which suggested that clustering would be most appropriate with only three clusters. A relatively small number of clusters is desirable in this setting because it is convenient, for example, to have small representative customer groups when designing customer tariffs. We also found that with a higher number of clusters, the clusters themselves became less meaningful.

4.2 Survey Classification

The unsupervised clustering of the smart meter time series allows us to assign a label to each smart meter, indicating the membership of each smart meter

to a electricity usage pattern clustering. These labels are then used to train a classification model in a supervised manner, to construct a mapping $h : \hat{Z} \to k$, where $\hat{Z} \equiv \{\hat{Z}_i\}_{i=1}^{3670}$, between the survey responses and the learned clusters. Constructing a mapping in this manner allows one to better understand the electricity usage patterns of a residential consumer using limited information about building characteristics. It is of interest to the electricity market to determine the most important of these features, so that targeted incentives and appropriate energy policies and climate plans can be designed.

Feature importance can be determined using wrapper methods, though these feature search methods can be computationally expensive if performed exhaustively. Instead, we perform our feature search using step backwards feature selection for the classification models. We perform the classification of the survey features using random forest classifiers and k-nearest neighbours classifiers, owing to the limited data available, their simplicity (and hence ease of interpretation), and in the case of the random forest models, so that we may also observe the feature importance values that are naturally computed during the learning process.

Classification Feature Selection Evaluating the feature importance using wrapper methods requires some level of care. Since multiple features can correspond to a one-hot encoding of the same survey question, and since we are interested in determining the most important survey question, we must take care to ensure that the backwards greedy feature selection process selects features by greedily searching through questions rather than elements of the survey vectors. This is achieved by creating a custom *scikit-learn* estimator to implement the fitting logic and using *mlxtend* to perform the wrapper method search. For each model we perform step backwards greedy feature selection, we use five-fold cross-validation and use ROC-AUC as the scoring measure.

5 Experimental Results

All experiments were carried out on a machine with 15.5 GB of RAM, with Ubuntu 18.04 and a six core Intel® Core™ i7-9750H CPU 2.60GHz processor. Each clustering and classification task was performed using tools from the *Scikit-Learn* Python package. Feature importance extraction was achieved using the *MLXtend* Python package. Due to limited time and a lack of code availability, it was not possible to make methodological comparison with the works of McLoughlin et al., Lavin and Klabjan, or Alonso et al. [1, 10, 12].

5.1 Feature Vectors

A variety of fixed-length feature vectors were constructed to test their usefulness for constructing clusters from the smart meter time series. The vector we denote by $\vec{d} \in \mathbb{R}^{48}$ contained 48 elements (corresponding to the 48 half-hours in a day), each representing the mean electricity consumption in kilowatt-hours for the

corresponding time-slot over the entire eighteen-month period of observation. That is, this vector represented the mean behaviour for a day over all recorded days. This vector was then normalised by dividing the value of each element by the sum of the values, so that the clustering would be agnostic of the magnitude of the electricity consumption. The vector denoted by $\vec{m} \in \mathbb{R}^{108}$ contains the vectors $\vec{d}_w \in \mathbb{R}^{48}$ and $\vec{d}_e \in \mathbb{R}^{48}$, which are the same as $\vec{d}$ but computed only over weekdays and weekend days respectively, along with a vector $\vec{n} \in \mathbb{R}^{12}$ representing the total energy usage for each month, normalised similarly. We also use the feature vector $\vec{w} \in \mathbb{R}^{336}$, which contains the mean value of electricity usage for each time-slot over an entire week, representing the "typical week". Finally, we also make comparison with the statistical feature vector $\vec{s} \in \mathbb{R}^{21}$ described in [5].

The survey data were normalised such that the maximum value that any element could take was unity and the minimum value was zero. The survey posed a respondent questions relating to the occupation, ages and number of residents in the house, whether they were present during the day, the age of the house, whether certain appliances were within it and how often they were used, as well as attitudes toward and expectations of the installation of the smart meter. For categorical features, such as the BER energy efficiency rating, a one-hot encoding was used. For discrete, ordinal features, their values were divided by the maximum possible value. In the case of the year of construction, this meant that the values were divided by 2009, the year that the study began, and re-scaled so that they took a minimum of zero. Such an assumption requires that new values falling outside this range much be clamped to the minimum and maximum values observed in this study. When values were unknown, they were imputed if imputation could be deemed reasonable. This resulted in a 110-dimensional vector, containing responses to the questions 200, 300, 420, 430, 43111, 4312, 4311, 4321, 4332, 433, 4352, 453, 6103, 460, 470, 4701, 471, 4801, 49002, 49004, 450, 452, 310, 401, 405, 410, and 4704. The statement of these questions and the permitted responses are given in Appendix A.

5.2 Behaviour Clusters

A number of algorithms were tested for clustering, but it was found that k-means with a ℓ_2 norm produced approximately equal-sized clusters reliably. We chose to partition the residential participants into three clusters, based on observations of cluster quality using the silhouette score. We performed the clustering on all 3670 feature vectors and obtain labels for each feature vector representation.

Figure 1 shows results for the clusters, the representative load profiles is the mean of the average daily electricity usage patterns for all members of clusters. We can see for the feature representations $\vec{d}$ and $\vec{w}$ that the produced clusters have approximately equal sizes and that the three clusters behave similarly in terms of average daily electricity usage. Differentiation between clusters is reflected in the usage curves, where one cluster exhibits strongly the expected diurnal electricity consumption pattern, where as the other shows much more consistent electricity usage throughout the day. Using the features $\vec{m}$ produces

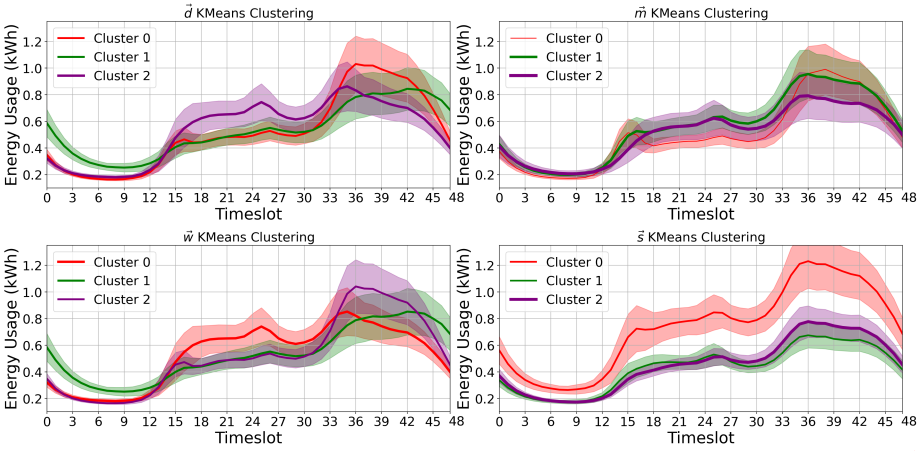

Fig. 1. Mean of the average daily electricity usage patterns for all members of clusters produced with a k means run. The shaded regions illustrate the variance of these mean values within the cluster and the thickness of the lines illustrate the relative sizes of the clusters, with the cluster having the most members represented by the thickest line.

two clusters of approximately equal size, and one smaller cluster. The behaviour of these clusters appears similar on average, but as we show later, we can establish membership of these more reliably from the information provided in the survey. The clusters produced from the features $\vec{s}$ demonstrate clusters that can be separated using consumption magnitude. Two of the clusters consume, in general, approximately equal magnitudes of electricity and illustrate some structural differences in their behaviour, however.

In order to assess the differentiating characteristics of each cluster, we analysed the survey responses associated with each meter. This was performed by determining a mean feature vector for each cluster and computing the variance between the mean question responses of different clusters, enabling us to identify the most discriminating questions. In Figure 2, we illustrate this by plotting the variance across the mean question response of different clusters, which indicates the discriminating potential of different questions. We observe that for all feature vector representations, the usage rates and ownership of specific appliances turn out to be important for characterising the membership of each cluster, as indicated by questions 49001 and 49004 (see Appendix A). Interestingly, one of the most important discriminating questions is question 405, asking if the household has access to the internet or not, suggesting that users with access to internet in 2009-2011 time period had a considerably different electricity usage pattern compared to those that didn't have internet access.

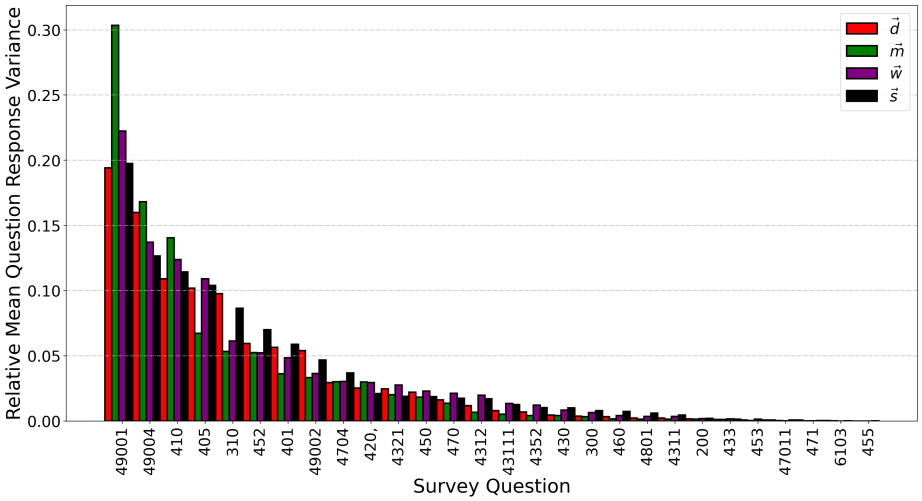

Fig. 2. Variance between the mean question response of different clusters, as a proxy to assess the discriminating potential of these survey questions in characterising the clusters – We consider the questions with high variance between their mean cluster responses as more discriminatory between clusters and the questions with similar answers across the different clusters as being less discriminatory.

5.3 Cluster Classification

Having been computed using normalised electricity usage vectors, the clusters produced are characterised by the attributes of the occupants of each building, and to a lesser extent the attributes of the building itself. This becomes further clear when we present the feature importance values based on the accompanying survey. It is possible to demonstrate which attributes these are by producing a histogram of survey responses for each cluster. In Figure 3, we can see that cluster 0 is much more likely to respond with option 1 for the employment question, indicating that they are employed, whereas clusters 1 and 2 have a large fraction of responses with option 6, indicating that they are retired. Similarly, we can see that cluster 1 is more likely to have one or two people over the age of 15 within the building during the day time, and more likely to have fewer bedrooms.

Inspecting the characteristics of those residential buildings that have been clustered shows that that the population is more likely to be distinguished by the composition of the occupants, the respondent's expectations and attitudes and the usage frequency of appliances within the residence than by the construction of the residence. In Figure 3, we see that for clusters produced from the features $\vec{m}$, the usage pattern corresponding to cluster one can be explained by the higher likelihood that it contains occupants who have reached pensionable age and who are less likely to have younger residents.

For the classification task, the cluster labels were used as supervised learning targets. Labels corresponding to the cluster embedding for each feature repre-

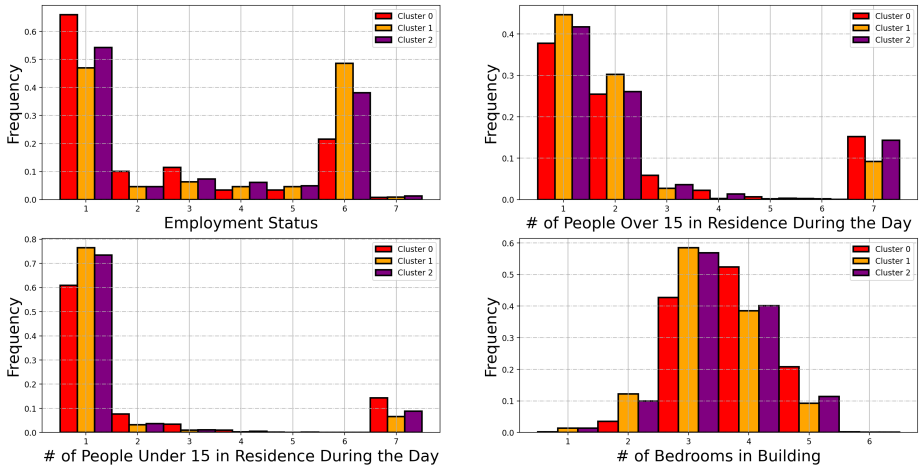

Fig. 3. Histograms illustrating the response frequencies for each cluster for select survey questions, where the clusters were constructed with the features $\vec{m}$.

sentation were tested, to determine which ones could be used to create clusters that facilitated the classification task well. The training data consisted of 75% of the participants, with the training and validation sets split evenly between the remaining 25% of the participants. The k-nearest neighbours model was tested for a variety of k values to determine the best values of $k \in \{1, 2, ..., 150\}$. The quality of each clustering model was determined using the testing and validation ROC-AUC and accuracy scores. In each case, the ℓ_2 norm was used as a measure of dissimilarity. Random forest models were constructed with between 100 and 1000 decision trees, using the information gain splitting technique. No maximum depth was specified and all other parameters were left as their default values according to the implementation in the *scikit-learn* package.

In Figure 4 we evaluate the ROC-AUC score for the k nearest neighbours models on the testing sets for a variety of values of k. In each case where a valid ROC-AUC score could not be computed, a point is omitted. In general, classification accuracy is relatively low, but can be improved for larger values of k, especially when computing clusters using the features $\vec{m}$. We note that this survey was not designed specifically for predicting the electricity usage patterns of the households and the relatively lower accuracy in our results is likely the result of the limited relevance of the survey questions to the underlying driving forces of electricity consumption profiles.

In Figure 5 we present the testing ROC-AUC scores for a variety of forest sizes. In several cases, the scores for the clusters generated using the statistical features are best, but this is unsurprising since variables corresponding to larger buildings will allow it to make distinctions more easily. We are not interested in magnitude profiling, so we ignore the statistical feature models when evaluating feature importance.

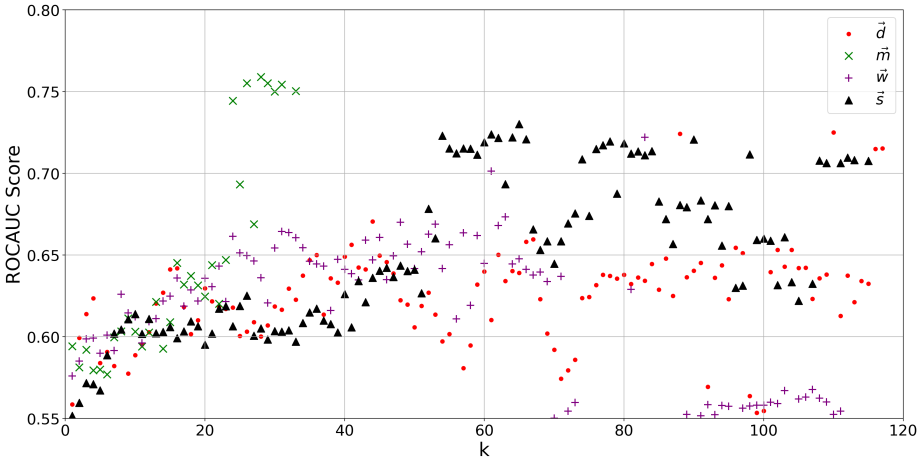

Fig. 4. ROC-AUC scores computed for k nearest neighbours classification for a range of values of k. Scores are computed for each of the feature representations.

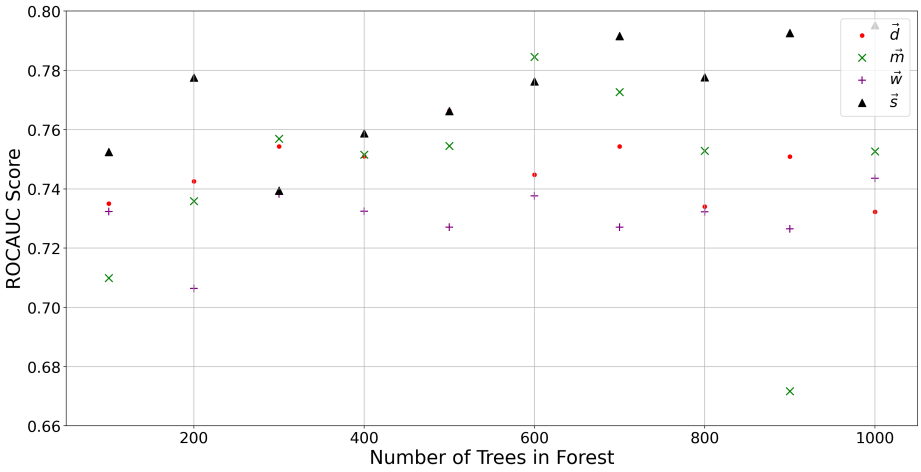

Fig. 5. ROC-AUC scores computed for the random forest classification models of various sizes. Scores are computed for each of the feature representations.

5.4 Feature Importance

Determination of the most important survey questions for correct classification of residential homes can be achieved by using a multitude of search algorithms, but performing this efficiently is difficult.

Evaluating the feature importance using wrapper methods requires some level of care. Since multiple features can correspond to a one-hot encoding of the same survey question, and since we are interested in determining the most important survey question, we must take care to ensure that the backwards greedy feature selection process selects features by greedily searching through questions rather than elements of the survey vectors. This is achieved by creating a custom *Scikit-Learn* estimator to implement the fitting logic and using *MLXtend* to perform the wrapper method search.

Feature importance values can be extracted from the random forest model implementation in *scikit-learn*. These indicate that the survey responses are dominated by few very important questions that translate to powerful features. Question 49004, is determined to be the most important, asking the respondent to indicate how often they use a variety of household appliances each day. Surprisingly, Question 4352, the next most important feature for the classification asks the participant how strongly they feel, either positively or negatively, that they can convince other occupants of the building to reduce their energy usage. The next three features included questions 49002, 49001 and 453 related to questions about how many entertainment devices of various kinds are in the home, how many household appliances of various types are in the home, and the year of construction. The most important single survey question was question 453. In Fig. 6 we can see that these five survey features remain the most important, irrespective of the features used to generate the clusters.

Performing backwards greedy feature selection for the variety of k-nearest neighbours models and random forest models outlined in the experiments above indicates that the features corresponding to these five questions are invariably the most important for classification accuracy. Although this has been computed for a limited spectrum of classification models, this suggests that these questions are, in general, the most important for classifying into the clusters constructed with relative electricity usage features.

6 Conclusions

In this work, we constructed a clustering of smart-meter time series for residential homes based individual average load profiles, deriving representative load profiles for the entire cluster. Using the cluster labels, we trained classification models to predict cluster membership using only occupancy, building construction and attitudinal survey responses. We identified the most relevant survey questions for performing such a classification, and those that are not, assigning relative importance values to each question obtained using random forest classifiers. We confirmed these results by performing step backwards greedy feature

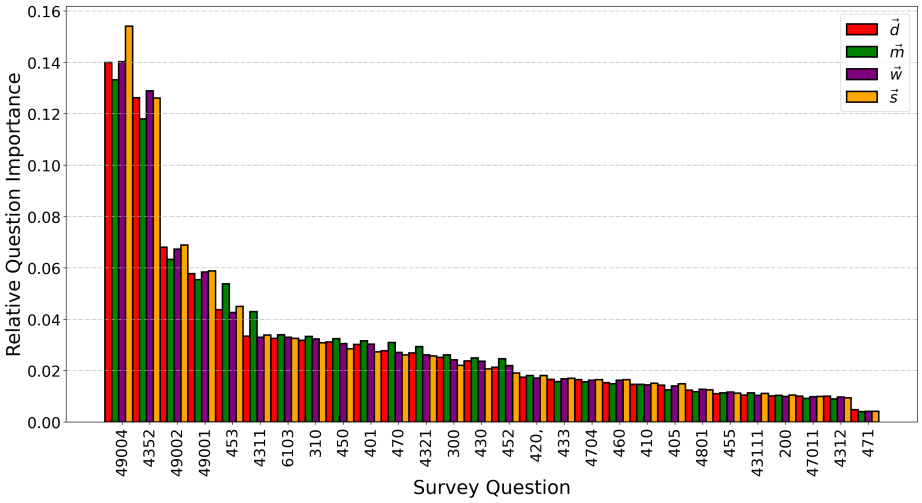

Fig. 6. Relative importance scores of survey questions, computed by the random forest classifiers. Importance values are computed for each case of the features used to determine the clusters.

selection, identifying usage of appliances, age of the building and attitudes of occupants towards energy usage as some of the most important characteristics to explain energy usage patterns. Unlike previous studies, we found that one of the most important characteristics of occupants of a residential household that influences their consumption behaviour is reflected by how likely it is that they feel they can convince other occupants to reduce their electricity consumption. The fact that a feature based on attitude of the people is more crucial to determining the electricity usage patterns compared to many other features based on characterizing the household and the building has important implications for policy makers, particularly in Ireland, where the returns on retro-fitting houses (as part of the climate action plan) has been found to be very poor. Our study suggests that a marketing campaign to alter the behavioural attitudes of people might be more effective in altering the usage patterns of residential customers.

It remains to determine precisely which questions would be more effective for improving the classification accuracy. Further work could be carried out to test alternative questions that will enable us to more accurately map the characteristics of a household to its energy usage patterns.

Acknowledgement

This work was funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183).

References

1. Alonso, A.M., Nogales, F.J., Ruiz, C.: Hierarchical clustering for smart meter electricity loads based on quantile autocovariances. IEEE Transactions on Smart Grid (2020)
2. Atluri, G., Karpatne, A., Kumar, V.: Spatio-temporal data mining: A survey of problems and methods. ACM Computing Survey **51**(4), 83:1–83:41 (2018)
3. Azaza, M., Wallin, F.: Smart meter data clustering using consumption indicators: responsibility factor and consumption variability. Energy Procedia **142**, 2236–2242 (2017)
4. Caroll, P., Dunne, J., Hanley, M., Murphy, T.: Exploration of electricity usage data from smart meters to investigate household composition. In: Conference of European Statisticians, 25-27 September 2013, Geneva, Switzerland (2013)
5. Carroll, P., Murphy, T., Hanley, M., Dempsey, D., Dunne, J.: Household classification using smart meter data. Journal of Official Statistics **34**(1), 1–25 (2018)
6. Commission for Energy Regulation (CER): CER smart metering project - electricity customer behaviour trial, 2009-2010. 1st Edition. Irish Social Science Data Archive. SN: 0012-00. www.ucd.ie/issda/CER-electricity (2012)
7. Esling, P., Agon, C.: Time-series data mining. ACM Computing Surveys (CSUR) **45**(1), 1–34 (2012)
8. Flath, C., Nicolay, D., Conte, T., van Dinther, C., Filipova-Neumann, L.: Cluster analysis of smart metering data. Business & Information Systems Engineering **4**(1), 31–39 (2012)
9. Ghasemkhani, A., Yang, L., Zhang, J.: Learning-based demand response for privacy-preserving users. IEEE Transactions on Industrial Informatics **15**(9), 4988–4998 (2019)
10. Lavin, A., Klabjan, D.: Clustering time-series energy data from smart meters. Energy efficiency **8**(4), 681–689 (2015)
11. Liao, T.W.: Clustering of time series data—a survey. Pattern recognition **38**(11), 1857–1874 (2005)
12. McLoughlin, F., Duffy, A., Conlon, M.: A clustering approach to domestic electricity load profile characterisation using smart metering data. Applied energy **141**, 190–199 (2015)
13. Meguelati, K., Fontez, B., Hilgert, N., Masseglia, F.: High dimensional data clustering by means of distributed dirichlet process mixture models. In: 2019 IEEE International Conference on Big Data (Big Data). pp. 890–899. IEEE (2019)
14. Motlagh, O., Berry, A., O'Neil, L.: Clustering of residential electricity customers using load time series. Applied Energy **237**, 11 – 24 (2019)
15. Mounce, S., Furnass, W., Goya, E., Hawkins, M., Boxall, J.: Clustering and classification of aggregated smart meter data to better understand how demand patterns relate to customer type. In: Proceedings of Computing and Control for the Water Industry (CCWI 2016) (2016)
16. O'Doherty, C.: Half of the homes in retrofit plan no better off despite cost. `https://www.independent.ie/irish-news/half-of-the-homes-in-retrofit-plan-no-better-off-despite-cost-39166389.html` (2020), article published on April 29 2020; accessed 16th June, 2020
17. Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E.J.: Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery **26**(2), 275–309 (2013)

A Survey Questions

Answers to the following questions were retained for use as features in the classification task. Note that questions 49003,1, 49003,2, 49003,3, 49003,4, 490004, 4900004, 4900005, 4900006, 4900007, 4900008, are encoded as question 49004 in the above results (that is, questions 46-55 below). Similarly, question 4551 is encoded as 455 in the results above.

1. 200 **PLEASE RECORD SEX FROM VOICE**
- □ Male
- □ Female

2. 300 **May I ask what age you were on your last birthday?**
- □ 18 - 25
- □ 26 - 35
- □ 36 - 45
- □ 46 - 55
- □ 56 - 65
- □ 65+
- □ Refused

3. 310 **What is the employment status of the chief income earner in your household, is he/she**
- □ An employee
- □ Self-employed (with employees)
- □ Self-employed (with no employees)
- □ Unemployed (actively seeking work)
- □ Unemployed (not actively seeking work)
- □ Retired
- □ Carer: Looking after relative or family

4. 401 **SOCIAL CLASS: Interviewer, Respondent said that occupation of chief income earner was....**
- □ AB
- □ C1
- □ C2
- □ DE
- □ F [RECORD ALL FARMERS]
- □ Refused
- □ Carer: Looking after relative or family

5. 410 **What best describes the people you live with?**
- □ I live alone
- □ All people in my home are over 15 years of age
- □ Both adults and children under 15 years of age live in my home

6. 420 **How many people over 15 years of age live in your home?**
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6
- ☐ 7 or more

7. 430 **And how many of these are typically in the house during the day (for example for 5-6 hours during the day)?**
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6
- ☐ 7 or more

8. 43111 **How many people under 15 years of age live in your home?**
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6
- ☐ 7 or more

9. 4312 **And how many of these are typically in the house during the day (for exanmple for 5-6 hours during the day)?**
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6
- ☐ 7 or more

10. 4331,3 **I / we am are interested in changing the way I / we use electricity if it reduces the bill**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

11. 4331,4 **I / we am are interested in changing the way I / we use electricity if it helps the environment**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

12. 4331,5 **I / we can reduce my electricity bill by changing the way the people I / we live with use electricity**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

13. 4321,2 **I / we have already done a lot to reduce the amount of electricity I / we use**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

14. 4321,3 **I / we have already made changes to the way I / we live my life in order to reduce the amount of electricity we use.**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

15. 4321,4 **I / we would like to do more to reduce electricity usage**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

16. 4321,5 **I / we know what I / we need to do in order to reduce electricity usage**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

17. 433 **Thinking about the energy reduction activities undertaken by you or your family/household, in the last year, did your efforts reduce your bills?**
- ☐ Yes
- ☐ No
- ☐ Don't know

18. 4352,2 **It is too inconvenient to reduce our usage of electricity**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

19. 4352,3 **I do not know enough about how much electricity different appliances use in order to reduce my usage**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

20. 4352,4 **I am not be able to get the people I live with to reduce their electricity usage**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

21. 4352,5 **I do not have enough time to reduce my electricity usage**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

22. 4352,6 **I do not want to be told how much electricity I can use**
- ☐ 1 - strongly agree
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 - strongly disagree

23. 4352,7 **Reducing my usage would not make enough of a difference to my bill**
- □ 1 - strongly agree
- □ 2
- □ 3
- □ 4
- □ 5 - strongly disagree

24. 450 **I would now like to ask some questions about your home. Which best describes your home?**
- □ Apartment
- □ Semi-detached house
- □ Detached house
- □ Terraced house
- □ Bungalow
- □ Refused

25. 452 **Do you own or rent your home?**
- □ Rent (from a private landlord)
- □ Rent (from a local authority)
- □ Own Outright (not mortgaged)
- □ Own with mortgage etc
- □ Other

26. 453 **What year was your house built**
- □ INT ENTER FOR EXAMPLE: 1981- CAPTURE THE FOUR DIGITS

27. 6103 **What is the approximate floor area of your home?**

28. 460 **How many bedrooms are there in your home?**
- □ 1
- □ 2
- □ 3
- □ 4
- □ 5 +
- □ Refused

29. 470 **Which of the following best describes how you heat your home?**
- □ Electricity (electric central heating storage heating)
- □ Electricity (plug in heaters)
- □ Gas
- □ Oil
- □ Solid fuel
- □ Renewable (e.g. solar)
- □ Other

30. 47001 **Do you have a timer to control when your heating comes on and goes off?**
- ☐ Yes
- ☐ No

31. 4701 **Which of the following best describes how you heat water in your home?**
- ☐ Central heating system
- ☐ Electric (immersion)
- ☐ Electric (instantaneous heater)
- ☐ Gas
- ☐ Oil
- ☐ Solid fuel boiler
- ☐ Renewable (e.g. solar)
- ☐ Other

32. 47011 **Do you have a timer to control when your hot water/immersion heater comes on and goes off?**
- ☐ Yes
- ☐ No

33. 4801 **Do you use your immersion when your heating is not switched on?**
- ☐ Yes
- ☐ No

34. 4704 **Which of the following best describes how you cook in your home**
- ☐ Electric cooker
- ☐ Gas cooker
- ☐ Oil fired cooker
- ☐ Solid fuel cooker (stove aga)

35. 471 **Returning to heating your home, in your opinion, is your home kept adequately warm?**
- ☐ Yes
- ☐ No

36. 49001,1 **Please indicate how many of the following appliances you have in your home? Washing machine**
- ☐ None
- ☐ 1
- ☐ 2
- ☐ More than 2

37. 49001,2 **Please indicate how many of the following appliances you have in your home? Tumble dryer**
- ☐ None
- ☐ 1
- ☐ 2
- ☐ More than 2

38. 49001,3 **Please indicate how many of the following appliances you have in your home? Dishwasher**
- ☐ None
- ☐ 1
- ☐ 2
- ☐ More than 2

39. 49001,4 **Please indicate how many of the following appliances you have in your home? Electric shower (instant)**
- ☐ None
- ☐ 1
- ☐ 2
- ☐ More than 2

40. 49001,5 **Please indicate how many of the following appliances you have in your home? Electric shower (electric pumped from hot tank)**
- ☐ None
- ☐ 1
- ☐ 2
- ☐ More than 2

41. 49001,6 **Please indicate how many of the following appliances you have in your home? Electric cooker**
- ☐ None
- ☐ 1
- ☐ 2
- ☐ More than 2

42. 49001,7 **Please indicate how many of the following appliances you have in your home? Electric heater (plug-in convector heaters)**
- ☐ None
- ☐ 1
- ☐ 2
- ☐ More than 2

43. 49001,8 **Please indicate how many of the following appliances you have in your home? Stand alone freezer**
- ☐ None
- ☐ 1
- ☐ 2
- ☐ More than 2

44. 49001,9 **Please indicate how many of the following appliances you have in your home? A water pump or electric well pump or pressurised water system**
- ☐ None
- ☐ 1
- ☐ 2
- ☐ More than 2

45. 49001,10 **Please indicate how many of the following appliances you have in your home? Immersion**
- ☐ None
- ☐ 1
- ☐ 2
- ☐ More than 2

46. 49003,2 **In a typical day, how often would you or your family/household use each appliance - please think of the total use by all household/family members. Washing machine**
- ☐ Less than 1 load a day typically
- ☐ 1 load typically
- ☐ 2 to 3 loads
- ☐ More than 3 loads

47. 49003,3 **In a typical day, how often would you or your family/household use each appliance - please think of the total use by all household/family members. Tumble dryer**
- ☐ Less than 1 load a day typically
- ☐ 1 load typically
- ☐ 2 to 3 loads
- ☐ More than 3 loads

48. 49003,4 **In a typical day, how often would you or your family/household use each appliance - please think of the total use by all household/family members. Dishwasher**
- ☐ Less than 1 load a day typically
- ☐ 1 load typically
- ☐ 2 to 3 loads
- ☐ More than 3 loads

49. 490004 **In a typical day, how often would you or your family/household use each appliance - please think of the total use by all household/family members. Electric shower (instant)**
- ☐ Less than 5 mins
- ☐ 5-10 mins
- ☐ 10-20 mins
- ☐ Over 20 mins

50. 4900004 **In a typical day, how often would you or your family/household use each appliance - please think of the total use by all household/family members. Electric shower (pumped from hot tank)**
- ☐ Less than 5 mins
- ☐ 5-10 mins
- ☐ 10-20 mins
- ☐ Over 20 mins

51. 4900004 **In a typical day, how often would you or your family/household use each appliance - please think of the total use by all household/family members. Electric shower (pumped from hot tank)**
- ☐ Less than 5 mins
- ☐ 5-10 mins
- ☐ 10-20 mins
- ☐ Over 20 mins

52. 4900005 **In a typical day, how often would you or your family/household use each appliance - please think of the total use by all household/family members. Electric cooker**
- ☐ Less than 30 mins
- ☐ 30-60 mins
- ☐ 1-2 hours
- ☐ Over 2 hours

53. 4900006 **In a typical day, how often would you or your family/household use each appliance - please think of the total use by all household/family members. Electric heater (plug-in)**
- ☐ Less than 30 mins
- ☐ 30-60 mins
- ☐ 1-2 hours
- ☐ Over 2 hours

54. 4900007 **In a typical day, how often would you or your family/household use each appliance - please think of the total use by all household/family members. Water pump**
- ☐ Less than 30 mins
- ☐ 30-60 mins
- ☐ 1-2 hours
- ☐ Over 2 hours

55. 4900008 **In a typical day, how often would you or your family/household use each appliance - please think of the total use by all household/family members. Immersion water**
- ☐ Less than 30 mins
- ☐ 30-60 mins
- ☐ 1-2 hours
- ☐ Over 2 hours

56. 4551 **What rating did your house achieve?**
- ☐ A
- ☐ B
- ☐ C
- ☐ D
- ☐ E
- ☐ F
- ☐ G

57. 405 **Do you have internet access in your home?**
- □ Yes
- □ No