What makes you peak? Cluster analysis of household activities and electricity demand

Aven Satre-Meloy
Environmental Change Institute
University of Oxford
3 S Parks Rd,
Oxford OX1 3QY
UK
aven.satremeloy@ouce.ox.ac.uk

Philipp Grünewald
Environmental Change Institute
University of Oxford
UK
philipp.grunewald@ouce.ox.ac.uk

Marina Diakonova
Environmental Change Institute
University of Oxford
UK
marina.diakonova@ouce.ox.ac.uk

Keywords

peak load, household electricity, cluster analysis, demand response, time use, demand patterns

Abstract

Researching the dynamics of energy consumption at finely resolved timescales is increasingly practical with the growing availability of high-resolution data and analytical methods to characterise them.

One of the methodological approaches that has recently become popular for exploring energy consumption dynamics is load profile clustering. Despite an abundance of available algorithmic techniques, clustering load profiles is challenging because clustering methods do not capture the temporal aspects of electricity consumption well and because cluster results are difficult to validate without detailed auxiliary data. These challenges make it difficult to use cluster analysis to better understand drivers of different electricity consumption patterns.

We address these challenges by applying a novel approach to a unique dataset of high-resolution electricity data, household socio-demographics, and occupant time-use data for a sample of 135 UK households.

Clusters can be identified for typical electricity use patterns and linked to activity patterns underlying these. We use this detailed data to validate load profile clusters, exploring how different socio-demographic data and patterns of household activity explain electricity consumption trends, focusing primarily on late afternoon and evening hours in the UK (4–9 pm), during which peak demand occurs.

We present peak-period clusters and the household characteristics and activities driving their demand. We show how such an approach can aid in segmenting classes of consumers to develop more targeted strategies for demand reduction and response interventions. This knowledge can be used to better understand the constraints and opportunities for a more flexible demand-side in the electricity sector.

Introduction

Reducing overall electricity demand in buildings through energy efficiency is an important component of climate mitigation strategies. But in addition to reducing total demand, better understanding the temporal aspects of electricity consumption is especially important in energy research. The domestic sector accounts for around 45 % of total UK electricity consumption and 50 % of UK national peak demand (BEIS, 2018; Ofgem, 2010). Reducing system peak demand lowers the costs and carbon-intensity of electricity generation. Delivering more responsive demand can also help integrate variable renewable energy into existing power systems. Beyond making electricity use more efficient, understanding how to increase the flexibility of use will benefit any low-carbon energy strategy (Grunewald and Diakonova, 2018).

Deeper insight into the factors that influence consumption patterns across hours of the day can inform solutions for demand flexibility and can improve efforts to target consumers for time-sensitive reductions in usage. Recent research suggests occupant activity data may be particularly valuable for understanding patterns of electricity consumption during different times of day (Satre-Meloy et al., 2018). Access to high-resolution electricity consumption and activity data can strengthen these insights if appropriate methods are used to characterise them. One technique used to identify temporal variations in electricity consumption that has gained popularity in recent years is cluster analysis of electric load profiles.

Cluster analysis is an algorithmic approach used to identify homogenous groupings of data where no a priori grouping exists. This data mining technique has gained popularity with the rise of 'big' data and machine learning, and it has been applied in recent research to interval meter data from residential customers (Rhodes et al., 2014; McLoughlin et al., 2015; Haben et al., 2016; Jin et al., 2017). This emerging research aims to apply various clustering algorithms to electricity load profiles for the purpose of identifying representative classes of residential customers, which can then be explained by socio-demographic or other household data. Such characterisation of load profiles can aid in understanding what drives variations in electricity use at finely-resolved timescales and can help identify targeted strategies for helping households shift demand.

Identifying robust links between household load profiles and other descriptive data faces two related challenges. First, there are numerous methodological considerations that must be addressed in order to ensure the clustering results are mathematically sound. These considerations are reviewed in the methods section of this paper. Second, because clustering is a technique that always yields some segmentation of data into groups, it can be difficult to gauge whether in addition to being mathematically sound, the clustering results are helpful. In other words, it is important to question whether clusters are useful for determining underlying patterns, such as the drivers of varying load profile patterns. Overcoming this challenge requires having access to various types of descriptive data to link to clusters.

This paper aims to improve upon previous residential load profile clustering efforts by introducing a novel approach for pre-processing and clustering household load profiles. We propose a simple approach to cluster cumulative rather than raw load profiles, which enables more appropriate use of Euclidean distance metrics to account for temporal rather than magnitude differences between clusters. This approach to pre-processing the data does not appear to have been used in previous cluster analyses of building load profiles. After pre-processing, we apply a hierarchical clustering algorithm to yield clusters that are more similar and distinct across their full shape than is the case when directly clustering raw load profiles. Using this approach results in household clusters that have clearly differentiated consumption trends from 4–9 pm, which is the time period we focus on for the purpose of understanding drivers of peak period electricity consumption.

This paper then performs an exploratory data analysis to link our unique dataset of household demographics, physical dwelling characteristics, appliance ownership, and time-use activity data to these clusters, showing how these data explain different peak-time user classes. The objective of this analysis is to show how access to household activity data, in addition to conventional demographic and appliance ownership variables, can improve segmentation of customers and identify the activities driving their demand. We show how such an approach can improve load profile clustering efforts for the specific purpose of developing more targeted and effective strategies for demand reduction and response interventions.

Methods

DATA COLLECTION AND SAMPLE

Data are collected as part of a five-year study for which data collection is ongoing (Grünewald and Layberry, 2015). Participating households, which are recruited online via e-mail and social media, complete a household survey before participating wherein they provide socio-demographic data along with physical dwelling characteristics and household appliance ownership details. Fuel type used for heating and cooking appliances is also collected here. Households then receive a parcel prior to their selected study date containing an electricity recorder, activity recorder(s), and an instruction booklet. Each household member over eight can participate.

On their selected date, which can be any day of the week excluding holidays, participants are instructed to attach the electricity recorder to their mains electricity. The recorder collects electricity consumption data at one-second resolution for 28 hours, from 5 pm until 9 pm on the following day. Around two-thirds of households participate on a weekday in the winter or shoulder seasons (September–April).

Activities are recorded using a purpose-built app that comes pre-installed on individual devices. The app guides participants through screens where they can enter the activity location, details of the activity, number of people participating, and enjoyment of the activity. The 'activity details' screens are tailored to collect numerous details on the type of activity, appliances used, and other energy-relevant details (see Figure 1). These screens enable participants to record more detail than would be possible using paper-based activity diaries, which have been used extensively for time-use research (eurostat, 2009). Participants are encouraged to record activities in real time, and activities are recorded at points in time rather than for durations in time. The app also provides for recording activities retrospectively and in the future. More details are discussed in Grünewald et al. (2017).

Several exclusion criteria are used to narrow down the study sample. Households where electricity records show continuous readings below 20 Watts are excluded (N=2), as this likely signals a failure to correctly attach the electricity recorder. Households that did not complete the full household survey are removed, and activities recorded outside the home are not included, since our primary interest is the direct relationship between electricity consumption and in-home activities (rather than those recorded at work, for instance). Finally, we restrict our data to the hours of 4–9 pm on the second day of the study and only include activities reported during these hours in order to investigate peak-time user clusters and activities driving their consumption patterns. Our final sample consists of 135 households and 350 individuals who together recorded 1,247 activities in the home during the 5-hour window from 4–9 pm.

DATA PRE-PROCESSING

Pre-processing data is an important part of any data analysis task and may include handling missing data and outliers, transforming and normalising data, and cleaning data prior to analysis. Previous research shows that clustering results are sensitive to numerous considerations with regard to 1) how to pre-process and prepare data for clustering, 2) how to define 'distance' for the clustering algorithm in order to group together data that are 'closer' or 'more similar', and 3) how to select an appropriate clustering algorithm. Jin et al. (2017) review these and other considerations in detail. Here, we discuss our approach.

Load profile normalisation

Normalisation is given close scrutiny in previous research on clustering load profiles. The goal of normalisation when clustering load profiles is driven by the analysis objective, which is typically either to cluster by magnitude of electricity consumption or by temporal variation in electricity use.

In past work where the aim was to segment consumers by the magnitude of their consumption, normalisation was done using a reference demand, such as the daily maximum demand of each household (Chicco, 2012; Chicco et al., 2006). In other research, usage was normalised by dividing by the range, often called min-max normalisation (Cao et al., 2013; Piao et al., 2014), or by the daily total (Kwac et al., 2014; Rhodes et al., 2014).

When the objective is segmentation by temporal variation in electricity use, or in other words, focusing on the relative magnitude of consumption in regard to the time of day, an alternative normalisation approach is more appropriate. In this paper, we apply a technique proposed by Jin et al. (2017) to normalise profiles such that clustering them will group together households with similar patterns of discretionary consumption at different times of day. The authors' method is to subtract daily minimum demand from hourly usage (called 'de-minning') and then divide each hour's consumption by the 'de-minned' total. Using this approach, daily minimum power demand is a proxy for baseload consumption, and after normalising the load profiles, "a load shape represents its hourly contribution to daily total discretionary usage and shape clusters can be interpreted in terms of timing of higher and lower discretionary demand" (Jin et al., 2017, p. 2). Before normalising, we aggregate electricity data and take the hourly average in order to yield smoother profiles and a simpler clustering process, but our method could be applied to electricity data measured at more frequent intervals. Figure 2 presents original hourly load profiles, and Figure 3 presents normalised load profiles for our full sample of households (N=135) during the study period (4–9 pm).

Distance metrics for cluster analysis

Since clustering works to group data objects that are similar into the same cluster and ensure these objects are dissimilar to those in other clusters, an important step in clustering is to determine how dissimilarity will be calculated. Determining dissimilarity is done using a distance measure, and many measures have been proposed in the clustering literature (Pandit and Gupta, 2011).

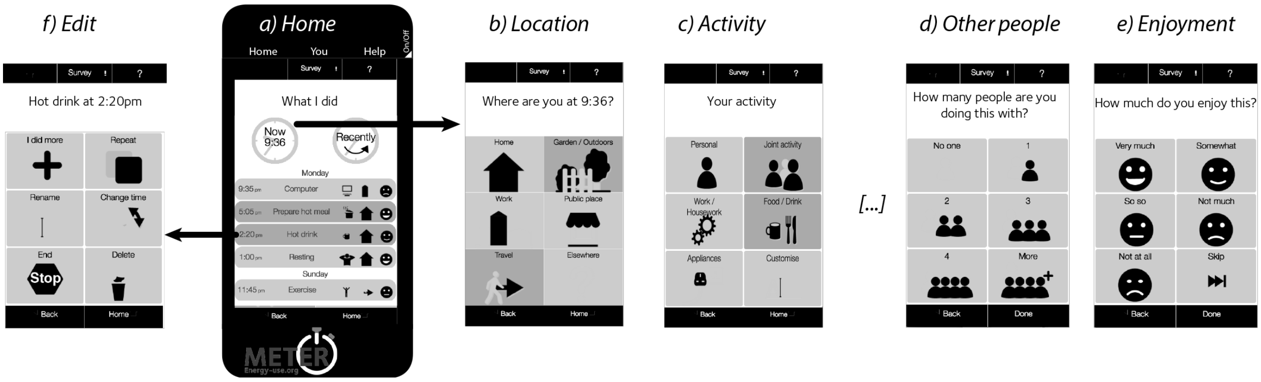

Figure 1. Activity-based time-use diary and sample activity entry sequence.

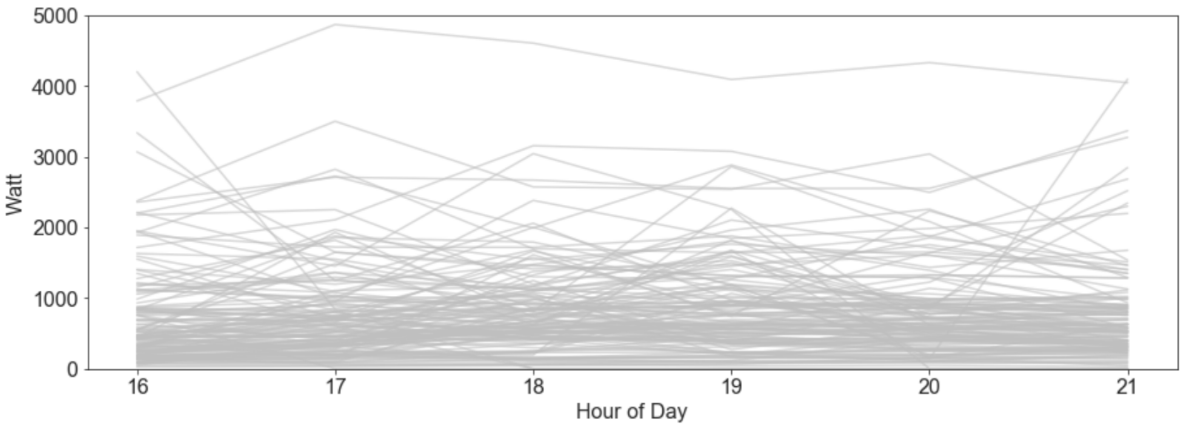

Figure 2. Hourly load profiles from 4–9 pm for full sample of households (N=135).

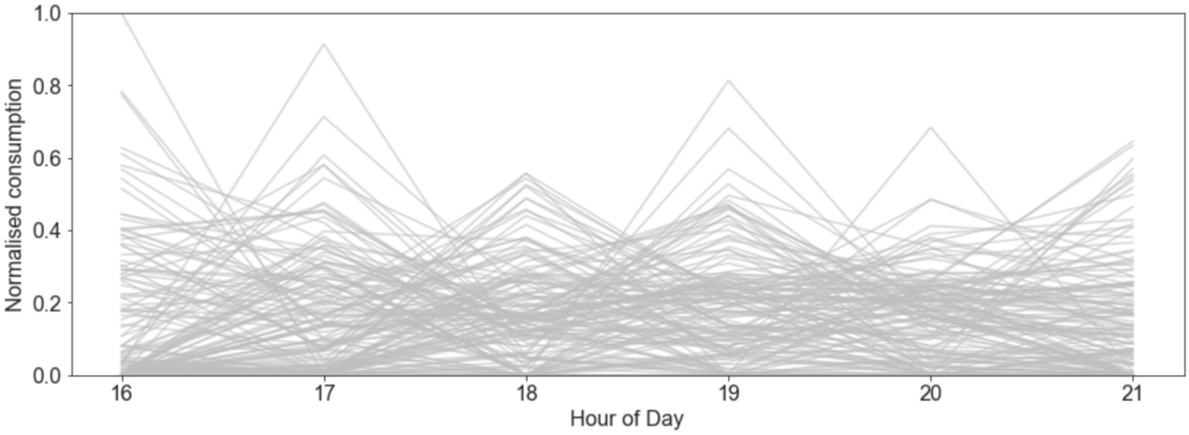

Figure 3. Normalised hourly load profiles from 4–9 pm for full sample.

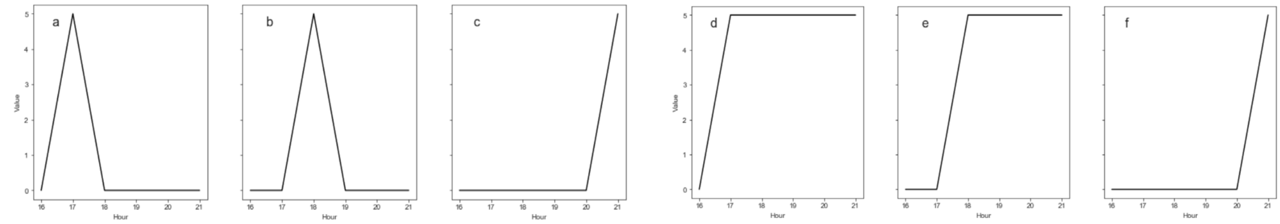

Figure 4. Example of three time series with equal Euclidean distances between each pair of untransformed time series (a–c) but with different Euclidean distances between each pair of cumulative time series (d–f).

The most well-known distance metric used for clustering continuous data is the Euclidean distance (Jain et al., 1999). The Euclidean distance is a useful metric for calculating the distance of objects in two or three-dimensional space. It is the straight-line distance between two points and is a variant of the Minkowski metric, which is defined as:

$$D(X,Y) = \sqrt[p]{\sum_{i=1}^{n} |(x_i - y_i)^p|} \qquad (1)$$

where $D(X,Y)$ gives the distance of order p between two points $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$. If $p = 2$, this equation gives the Euclidean distance, and if $p = 1$ it gives the Manhattan or city block distance, which is the distance between two points measured along axes at right angles and is so-named because it is achieved by walking 'around the block' compared to the Euclidean 'straight line' distance. If $p = \infty$, the equation gives the Chebyshev distance, which is the greatest of the differences between two vectors along any coordinate dimension (Chicco, 2012).

Several other variants of Euclidean distance include the squared Euclidean distance (sqeuclidean), which places greater weight on objects that are farther apart, and standardized Euclidean distance (seuclidean), which divides the Euclidean distance by the variance of all the i^{th} components of the objects, thus achieving a similar aim as normalising data prior to measuring distance.

Additional distance metrics intended for use with continuous data, and which have been proposed for clustering load profiles, include cosine dissimilarity, which measures the cosine of the angle between two vectors, and correlation dissimilarity, which uses the correlation coefficient as the dissimilarity indicator.

Importantly, these metrics are used when clustering data in Euclidean space. Clustering time series, such as hourly load profile data, introduces challenges because traditional distance metrics such as those described above are inadequate for determining dissimilarity. This is because Euclidean distances do not capture the ordered, temporal aspect of time-series data. A simple example of this is given in Figure 4, which shows three 6-element vectors with magnitude 5 plotted both as untransformed (*a–c*) and cumulative (*d–f*) time series. Using the Minkowski equation with $p = 2$, the Euclidean pairwise distance for each pair of the untransformed time-series plots in Figure 4 (*a–c*) is equal to $\sqrt{50}$. In other words, clustering these time series using Euclidean distance would place these points equidistant from each other.

We propose a simple fix to this issue. To the authors' knowledge, this approach has not been used previously to pre-process load profiles prior to clustering. Instead of clustering untransformed time series, we first integrate with respect to time and cluster cumulative time series. The effect of doing so is shown in plots *d–f* of Figure 4. Now, the Euclidean distance between *d* and *e* is $\sqrt{25}$, the distance between *d* and *f* is $\sqrt{100}$, and the distance between *e* and *f* is $\sqrt{75}$. Clustering these cumulative time series would group together *d* and *e*, which we argue is more appropriate given that they occur closer together in time.

This approach avoids the aforementioned issues and permits the appropriate use of Euclidean distances and their variants for clustering load profile data. While other approaches have been proposed for clustering time series data (e.g. Teeraratkul et al. (2018)), we argue our approach is advantageous both in its conceptual simplicity as well as its ease of implementation. In combination with the normalisation approach described previously, we cluster cumulative, normalised hourly load profiles, thus achieving the simultaneous goals of focusing on temporal variations in discretionary usage while also enabling straightforward Euclidean distance clustering. Figure 5 presents the cumulative, normalised hourly load profiles for the full sample during the 4–9 pm period.

Choice of clustering algorithm

A final consideration is the choice of clustering algorithm. A wide range of conventional and novel clustering methods have been applied to residential load profiles in the existing literature, and these generally fall under three families of clustering techniques: centroid-based methods, which group each observation to its closest centroid and iteratively update centroids until convergence (when the centroids no longer change); hierarchical clustering methods, which use a linkage criterion to determine the distance between different clusters and then successively merge these clusters in an agglomerative fashion; and model-based clustering, which fits a probability distribution to each cluster, or component, and assigns data points based on the probability of belonging to each cluster. See Jin et al. (2017) for an overview of these methods.

We use a hierarchical agglomerative clustering (HAC) algorithm rather than a centroid- or model-based method. HAC algorithms start with an initial cluster assignment where each observation (household) forms its own singleton cluster. Then, using a chosen distance metric to compute a dissimilarity matrix for the data, the singleton clusters are grouped into binary clusters using a linkage criterion. This process is repeated in an iterative fashion, merging clusters at each successive level until all observations are grouped into a single cluster (Chicco et al., 2006).

There are several linkage criteria that can be used with an HAC algorithm. These vary slightly in how they measure the distance between clusters at each successive level. ‘Single linkage’ measures the distance between the closest pairs of observations between two clusters, while ‘complete linkage’ measures the distance between the furthest pairs of observations. The ‘average linkage’ criterion measures the average distance between all cluster members, which can avoid the tendency of single linkage to form few large clusters or the tendency of complete linkage to form many smaller clusters. Two additional linkage criteria we consider are ‘Ward’, which forms clusters by minimising the within-cluster sum of squares at each iteration, and ‘centroid’, in which the similarity of two clusters is defined by the similarity of their centroids.

We apply an HAC algorithm instead of other methods for several reasons. First, in contrast to other methods, HAC algorithms do not require advance knowledge of the number of clusters and output an intuitive, visual representation of the clustering process. This visual is a dendrogram, or tree diagram, and it shows how clusters are merged at successive steps and thus gives a visual history of the clustering process. This can aid in determining the number of clusters by signalling where to ‘cut’ the tree, which is typically done where a large increase in distance between existing clusters is found. HAC is also advantageous in that it is easy and straightforward to implement. The disadvantages of HAC are that it can be computationally expensive when applied to large datasets, it does not handle missing data well, and it works poorly with mixed data types. Given the relatively small dataset, the absence of missing data, and the single data type used to cluster households in the present study, we use HAC for its simplicity and ease of interpretation.

CLUSTER AND EXPLORATORY DATA ANALYSES

After normalising and transforming raw hourly load profiles for our sample of households, we apply hierarchical clustering to these profiles using a combination of different Euclidean distance metrics and linkage criteria. For each distance-linkage combination, we compute the cophenetic correlation coefficient (CCC), which measures the linear correlation between pairwise distances for original observations and modelled dendrogram distances for those observations (Jin et al., 2017). In other words, the CCC measures how well the HAC dendrogram preserves distances between original data points. A value of 1 indicates perfect correlation. Figure 6 presents a heatmap of the CCC values for each distance-linkage combination. The heatmap shows that average and complete linkage

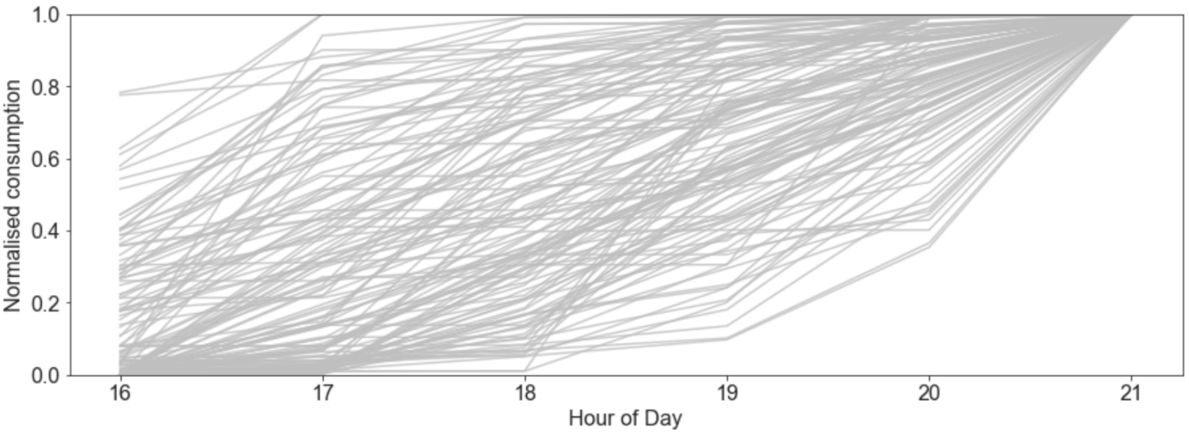

Figure 5. Cumulative normalised hourly load profiles from 4–9 pm for full sample.

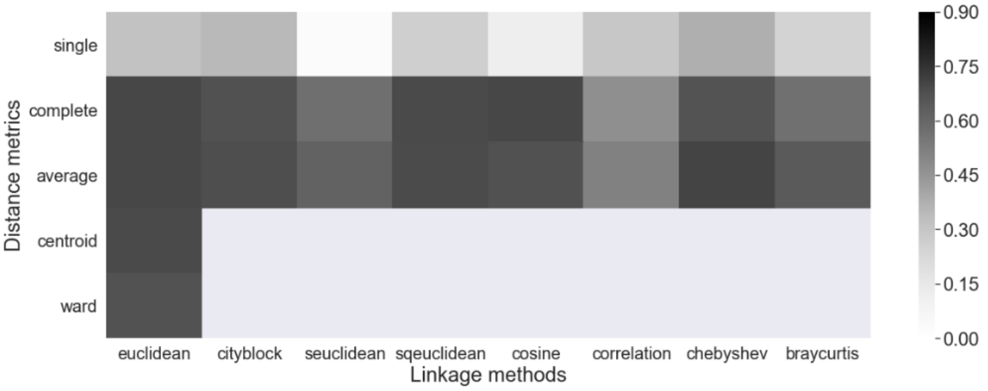

Figure 6. Heatmap of CCC values across combinations of distance and linkage methods used in hierarchical clustering.

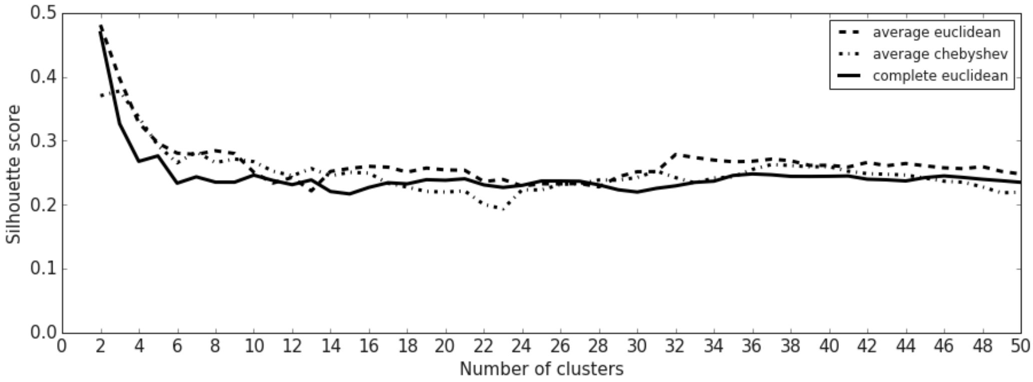

Figure 7. Silhouette scores for varying numbers of clusters across three distance metric and linkage criterion combinations.

combinations outperform single linkage combinations, and it shows that standardised Euclidean and correlation dissimilarity metrics do not preserve the original data structure as well as the other metrics. We select the top three combinations and compute dendrograms for each. These are average linkage with Chebyshev distance, complete linkage with Euclidean distance, and average linkage with Euclidean distance.

To perform final clustering, we use a clustering validity indicator, the silhouette index, to determine the optimal number of clusters across each of the three selected distance-linkage combinations. There are many different cluster validity indicators reported in the clustering literature, and these are used to assess how 'compact' and 'distinct' the final clusters are. The silhouette index is used to measure both compactness and distinctness simultaneously (Rousseeuw, 1987). The silhouette score is given by:

$$Sil_{score} = \frac{\max{(a, b)}}{b - a} \qquad (2)$$

where a is the average intra-cluster distance and b is the average shortest distance to another cluster. The silhouette score ranges from [-1, 1], where a high value indicates the observations in the cluster are well matched to each other and poorly matched to observations in other clusters. Computing the silhouette scores across a range of numbers of clusters (in other words, a range of values at which to cut the dendrogram) enables comparison of the validity of each distance-linkage approach for varying numbers of clusters. Figure 7 shows how the silhouette scores vary for each of the three chosen combinations of distance metrics and linkage criteria. This figure exhibits a typical elbow shape, indicative of scores that are high for small numbers of clusters but decrease and stay low as the number of clusters increases. Scores are relatively similar across distance-linkage combinations, so choosing the method and number of clusters is also guided by the objective of yielding clusters with more even distributions in cluster size.

We investigate dendrograms to make a final selection. The method that gives the most even cluster sizes while also maintaining a silhouette score and CCC that compare well to the other methods is the combination of Euclidean distance with the complete linkage criterion and three clusters. These are the parameters that are used to cluster the cumulative, normalised hourly load profiles for our sample of households, and these are the clusters on which our subsequent exploratory analyses are based. In the following section, we present cluster results along with how clusters are differentiated by socio-demographic, physical dwelling, appliance ownership, and activity pattern data.

Results

HIERARCHICAL CLUSTERING RESULTS

We apply HAC to cumulative, normalised hourly load profiles for our sample of 135 households during the hours of 4–9 pm. We use Euclidean distance and the complete linkage criterion to cluster load profiles. Cluster results are validated using the silhouette index, and the number of clusters is set to three to balance selecting a cluster assignment with a comparatively high silhouette score as well as a more even distribution of cluster sizes.

Figure 8 shows cluster results plotted with all cumulative, normalised load profiles, and Figure 9 shows cluster averages plotted with original, non-transformed load profiles. These plots show how clusters are differentiated in terms of their peak-period usage profiles. Cluster 1, which includes 37 households, has an early peak around 4 pm and has decreasing hourly usage through the rest of the evening. Cluster 2, the largest cluster with 75 households, shows increasing hourly consumption until its peak around 7 pm and then steadily decreasing consumption. Cluster 3 is the smallest cluster with 23 households, and this cluster has the lowest consumption in the early afternoon with steadily increasing consumption through the evening and a late peak at 9 pm. The peak totals for each cluster are similar, with cluster 1's peak total equal to 1,208 W at 4 pm, cluster 2's peak total equal to 1,030 W at 7 pm, and cluster 3's peak total equal to 1,050 W at 9 pm. During the typical time for UK system peak demand, the group with the highest demand are the 'early peak' households in cluster 1. These therefore deserve particular attention in efforts to reduce peak demand (DECC, 2014).

SOCIO-DEMOGRAPHIC, DWELLING, AND APPLIANCE OWNERSHIP CHARACTERISTICS

Clusters are investigated for differences in socio-demographic, dwelling, and appliance ownership characteristics to explore how these factors might explain peak-period usage profiles. Table 1 presents descriptive statistics for each of the survey variables across all three clusters and for the full sample. Several differences stand out between clusters.

Cluster 1 has a lower frequency of flats/apartments but a higher frequency of terraced houses than the other clusters and full sample. It also has the highest frequency of homeowners (as opposed to renters) and the highest mean number of occupants over 50 years old. Conversely, it also has the highest mean number of young occupants (under the age of 18). This cluster is also characterised by higher ownership of washing machines and tumbler dryers, TV/computer screens, night storage heaters, and power showers than the other clusters.

The households in cluster 2 are more likely to be detached or semi-detached, are primarily owned rather than rented, and more often have electric hobs and portable heaters than the other clusters. This cluster also exhibits the highest frequency of households with annual income over £50,000 and 1–2 occupants, along with the lowest mean number of occupants under 18. This cluster is most similar to the full sample, which

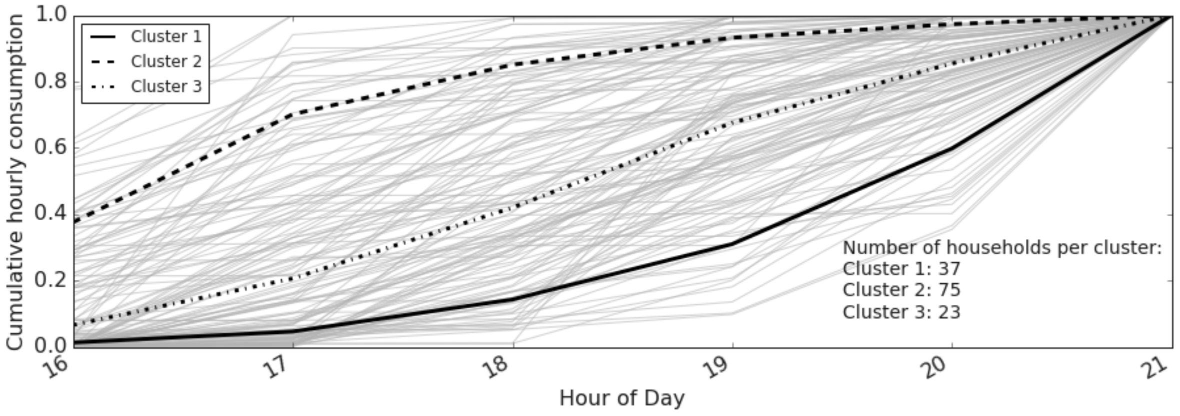

Figure 8. Cumulative, normalised hourly load profiles with cluster averages plotted for each cluster.

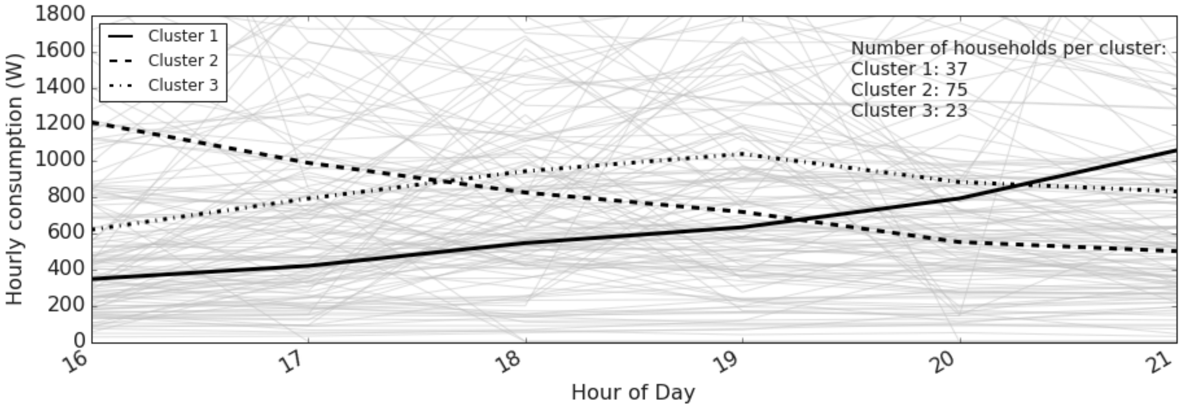

Figure 9. Non-normalised, original load profiles plotted with cluster averages for each cluster.

Table 1. Descriptive statistics for select socio-demographic, dwelling, and appliance ownership variables for each cluster and for full sample. Mean (M) and standard deviation (SD) are given for continuous variables.

Description	Response	Cluster 1 (4 pm peak) (n=37) %	Cluster 2 (7 pm peak) (n=75) %	Cluster 3 (9 pm peak) (n=23) %	Full sample (n=135) %
Home type	Flat/apartment	16	21	41	21
	Detached	24	27	13	24
	Semi-detached	22	28	26	26
	Terraced and other	38	24	30	29
Tenure	Rent	14	16	26	17
	Own	86	84	74	83
No. of rooms	2 or fewer	5	5	17	7
	3–5	49	42	52	53
	6 or more	46	53	31	40
Household income	<£25,000	22	16	22	19
	<£35,000	5	11	9	9
	<£50,000	22	19	22	20
	>£50,000	51	54	48	52
No. of occupants	1–2	49	63	47	56
	3–4	51	37	53	44
No. of occupants by age (count)	Under 18	*M = 0.99, SD = 1.77*	*M = 0.61, SD = 1.41*	*M = 0.78, SD = 1.37*	*M = 0.75, SD = 1.5*
	19–50	*M = 1.16, SD = 1.64*	*M = 1.31, SD = 1.69*	*M = 1.56, SD = 1.86*	*M = 1.26, SD = 1.65*
	Over 50	*M = 0.78, SD = 1.41*	*M = 0.61, SD = 1.15*	*M = 0.35, SD = 0.65*	*M = 0.61, SD = 1.20*
Large appliances	Washing machine	97	83	74	85
	Tumble dryer	46	28	35	34
	Washer dryer	8	19	22	16
	Gas boiler	89	76	70	79
	Heat pump	5	4	4	4
	Electric hob	24	40	35	35
Other appliances (count)	TV/computer screens	*M = 3.11, SD = 1.59*	*M = 2.65, SD = 1.74*	*M = 2.17, SD = 1.19*	*M = 2.7, SD = 1.64*
	Air conditioners	*M = 0.08, SD = 0.36*	*M = 0.03, SD = 0.16*	*M = 0.09, SD = 0.29*	*M = 0.05, SD = 0.25*
	Portable heaters	*M = 0.35, SD = 0.72*	*M = 0.51, SD = 0.72*	*M = 0.39, SD = 0.66*	*M = 0.44, SD = 0.71*
	Night storage heaters	*M = 0.27, SD = 1.04*	*M = 0.07, SD = 0.30*	*M = 0.17, SD = 0.84*	*M = 0.14, SD = 0.68*
	Power showers	*M = 0.41, SD = 0.69*	*M = 0.32, SD = 0.64*	*M = 0.22, SD = 0.52*	*M = 0.33, SD = 0.63*

could be expected because it accounts for more than 50 % of the full sample.

Cluster 3 contains the highest frequency of flats/apartments and renters, the lowest mean number of occupants aged over 50, and somewhat lower large appliance ownership than the other two clusters. There is also a lower frequency of large dwellings in this sample and of households with annual income over £50,000.

PEAK-TIME ACTIVITY PATTERNS

Activities reported from 4–9 pm by each cluster of households are grouped and investigated to understand how differences in the frequency of activities and their timing might explain differences in electricity consumption patterns. For this exploratory analysis, activity histograms are created to display the distribution of activity times for each cluster of households. We focus on activities that are reported most frequently across clusters as well as those we expect have some bearing on timing of peak usage, such as activities involving large appliances. These activity histograms are presented in Figure 10. Each activity histogram presents frequency counts for the median time an activity is reported by households in each cluster. We plot the median activity time for the entire household rather than the timing of activities recorded by individuals in each household for simplicity and to ensure the activity time for each household is not affected by 'outlying' individuals within each household. Some information on activities may be lost here, but for the purpose of investigating cluster-wide timing of different activities, we believe the median is a suitable metric. Each plot in Figure 10 also displays the mean and median timing of the activity for the whole cluster, which enables a comparison between them.

Activities that have the highest frequency across the full sample are eating a hot meal, watching TV, using the kettle or having a hot drink, 'custom' activity, cooking with the oven or on the hob, reading, arriving home, socialising, and computer/ Internet time. Watching TV is the most frequent activity for cluster 1, while reading is the most frequent for cluster 3.

In terms of activity timing, Figure 10 shows that activity patterns between clusters are not, on the whole, clearly differentiated. Food-related activities do, however, show some bearing on timing of peak usage, especially the timing of cooking on the hob or using the oven. The cluster with an earlier peak (4 pm) has a mean time for this activity that is a full hour before the '7 pm-peak' cluster, which in turn reports a mean cooking time that is 30 minutes earlier than the '9 pm-peak' cluster. The timing of the evening meal for the '4 pm-peak' cluster is nearly an hour earlier than in the '7 pm-peak' cluster and half an hour earlier than in the '9 pm-peak' cluster.

Results for recreational activities are mixed. The mean timing of watching TV for each cluster is within a period of 30 minutes (19:06–19:29). Mean computer and Internet use occurs around the same time in the '4 pm-peak' and '7 pm-peak' clusters (18:44 and 18:28, respectively), but slightly later in the '9 pm-peak' cluster (19:04).

Though they are not shown in Figure 10, several other activity histograms were investigated. Mean evening shower time occurs earlier in the '4 pm-peak' cluster than in the other clusters, and the meantime for socialising occurs about an hour later in the '9 pm-peak' cluster than in the other two. Mean timing of washing dishes occurs later in the '4 pm-peak' cluster (19:59) and '9 pm-peak' cluster (19:34) than in the '7 pm-peak' cluster (18:57).

Discussion and conclusions

This paper presents an integrated cluster and exploratory data analysis of hourly peak-period electricity consumption for 135 UK households. It addresses key challenges inherent to clustering load profiles through several important data pre-processing steps. First, we normalise load profiles using an approach such that the timing of discretionary usage is the main criteria on which households are clustered. Second, we integrate household load profiles with respect to time in order to more appropriately use conventional clustering metrics while focusing on the temporal variations in consumption. The results of our cluster analysis show three representative usage classes with relatively even sizes. Clusters include one that peaks in the early afternoon at 4pm and decreases its usage through the evening, one that increases until its peak at 7 pm and then decreases thereafter, and one that slowly increases through the evening until a late peak at 9 pm.

Next, our exploratory analysis shows that while socio-demographic and activity data do not completely differentiate clusters of households, they do indicate some patterns underlying peak-time usage patterns. The early-peaking cluster, which is the largest contributor of the three clusters to system peak demand (which occurs at 5 pm), includes households with older occupants, more large appliances, and fewer renters. The late-peaking cluster has the highest fraction of renters and individuals living in flats or apartments as well as fewer older occupants, large dwellings, and large appliances. The cluster with a peak between these two has a larger share of high-income households and a lower share of young occupants.

Patterns in timing of afternoon and evening activities also show some bearing on varying peak times and usage characteristics. The strongest link is found for food preparation and eating activities, as energy-intensive activities such as cooking with the oven or hob are more clearly differentiated across clusters than are recreational activities, such as watching TV or using the computer.

These results have implications both for methodological approaches to electricity consumption characterisation as well as our understanding of the drivers of residential peak demand. The methodological implications are that cluster results are highly sensitive to data pre-processing techniques. While normalising load profiles prior to clustering is common in most studies, considerations of which distance metrics to use with time-series data are less often discussed. We show a simple technique to appropriately use Euclidean distances with time-series data.

The empirical implications of our results show how occupant socio-demographics and activities relate to the temporal aspects of household electricity consumption. Usage curtailment potential varies considerably across customers, so understanding the factors that make customers more or less able to shift or curtail demand at different times of day can help to achieve greater responses. For instance, our results suggest that the households that contribute the most to UK system-wide peak demand are those with older occupants that have earlier cooking and meal times. This knowledge is useful for considering appropriate interventions, as evening meal practices may be particularly challenging to shift. Our results related to household income are mixed, but the late peaking cluster includes more lower-income households and also more renters. Design of targeted interventions, especially those relying on a price signal to shift demand, will benefit from taking this knowledge into account. Lower-income households as well as renters may have less ability to respond and thus may be more adversely impacted by time-of-use tariffs than homeowners and wealthier households.

Some limitations are present in the data and analysis. The sample size for this study is small, and cluster results are thus more prone to error. Our sample also consists of study days that vary across seasons and days of the week (though most of our sample participated on a winter weekday). Day of week and seasonal differences are no doubt important for understanding the temporal aspects of activities and electricity consumption. We also expect biases are present in our sample's activity reporting and note biases in its socio-demographic makeup, such as an overrepresentation of high-income and well-educated households. Participant self-selection is a further source of bias.

In terms of analysis, only one cluster validity index is used to determine the optimal number of clusters, and because we also take into account evenness of cluster sizes, this decision is ultimately a subjective one. Use of additional indices or comparison with other clustering algorithms would make this analysis more robust. Our analyses of socio-demographic, dwelling, and activity differences between clusters are exploratory in nature, and we plan in future to conduct full statistical analyses of these relationships using formal modelling techniques.

Accepting these limitations, the approach we demonstrate can help identify constraints and opportunities in achieving more responsive demand in the residential sector. We expect this approach to yield greater insights when applied to a larger,

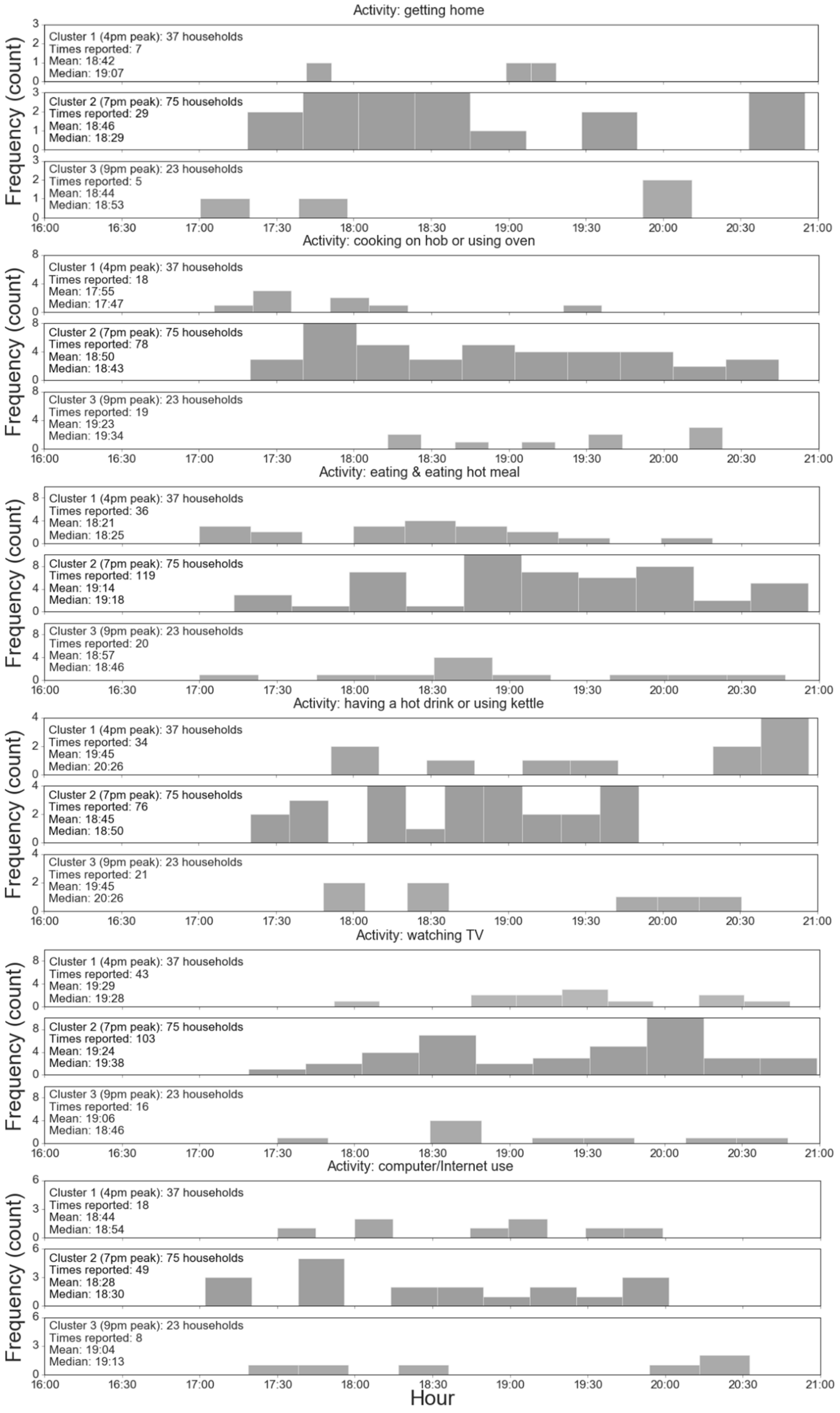

Figure 10. Histograms of median activity time for six frequently-reported activities across load profile clusters.

more representative sample. Data collection is on-going. There remain substantial opportunities to improve the effectiveness of demand response interventions, and we encourage more detailed investigations of high-resolution electricity data and occupant activities to ensure these deliver energy and cost saving benefits without adversely affecting vulnerable populations.

References

BEIS, 2018. DUKES chapter 5: statistics on electricity from generation through to sales, in: Digest of UK Energy Statistics (DUKES): Electricity. Department for Business, Energy & Industrial Strategy, London, United Kingdom.

Cao, H.-A., Beckel, C., Staake, T., 2013. Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns. Presented at the IECON 2013–39th Annual Conference of the IEEE Industrial Electronics Society, IEEE, Vienna, Austria, pp. 4733–4738. https://doi.org/10.1109/IECON.2013.6699900

Chicco, G., 2012. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 42, 68–80. https://doi.org/10.1016/j.energy.2011.12.031

Chicco, G., Napoli, R., Piglione, F.P., 2006. Comparisons Among Clustering Techniques for Electricity Customer Classification. IEEE Transactions on Power Systems 21, 933–940. https://doi.org/10.1109/TPWRS.2006.873122

DECC, 2014. Seasonal variations in electricity demand. Department of Energy and Climate Change, London, United Kingdom.

eurostat, 2009. Harmonised european time use surveys: 2008 guidelines. Office for Official Publications of the European Communities, Luxembourg.

Grunewald, P., Diakonova, M., 2018. Flexibility, dynamism and diversity in energy supply and demand: A critical review. Energy Research & Social Science 38, 58–66. https://doi.org/10.1016/j.erss.2018.01.014

Grünewald, P., Diakonova, M., Zilli, D., Bernard, J., Matousek, A., 2017. What we do matters – a time-use app to capture energy relevant activities, in: Proceedings of the 2017 eceee Summer Study on Energy Efficiency in Buildings. eceee, Presqu'île de Giens, France, pp. 2085–93.

Grünewald, P., Layberry, R., 2015. Measuring the relationship between time-use and electricity consumption, in: Proceedings of the 2015 eceee Summer Study on Energy Efficiency in Buildings. eceee, Presqu'île de Giens, France, pp. 2087–96.

Haben, S., Singleton, C., Grindrod, P., 2016. Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data. IEEE Transactions on Smart Grid 7, 136–144. https://doi.org/10.1109/TSG.2015.2409786

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Computing Surveys 31, 264–323. https://doi.org/10.1145/331499.331504

Jin, L., Lee, D., Sim, A., Borgeson, S., Wu, K., Spurlock, C.A., Todd, A., 2017. Comparison of Clustering Techniques for Residential Energy Behavior using Smart Meter Data. Presented at the Thirty-First AAAI Conference on Artifical Intelligence, p. 7.

Kwac, J., Flora, J., Rajagopal, R., 2014. Household Energy Consumption Segmentation Using Hourly Data. IEEE Transactions on Smart Grid 5, 420–430. https://doi.org/10.1109/TSG.2013.2278477

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. https://doi.org/10.1016/j.apenergy.2014.12.039

Ofgem, 2010. Demand Side Reponse. A Discussion Paper. Office of Gas and Electricity Markets, London.

Pandit, S., Gupta, S., 2011. A Comparative Study on Distance Measuring Approaches for Clustering. International Journal of Research in Computer Science 2, 29–31. https://doi.org/10.7815/ijorcs.21.2011.011

Piao, M., Shon, H.S., Lee, J.Y., Ryu, K.H., 2014. Subspace Projection Method Based Clustering Analysis in Load Profiling. IEEE Transactions on Power Systems 29, 2628–2635. https://doi.org/10.1109/TPWRS.2014.2309697

Rhodes, J.D., Cole, W.J., Upshaw, C.R., Edgar, T.F., Webber, M.E., 2014. Clustering analysis of residential electricity demand profiles. Applied Energy 135, 461–471. https://doi.org/10.1016/j.apenergy.2014.08.111

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Satre-Meloy, A., Diakonova, M., Grünewald, P., 2018. Daily Life and Demand: New Data on Behavioral Drivers of Residential Electricity Use Patterns, in: Proceedings of the 2018 ACEEE Summer Study on Energy Efficiency in Buildings. ACEEE, Asilomar, CA, p. 12.

Teeraratkul, T., O'Neill, D., Lall, S., 2018. Shape-Based Approach to Household Electric Load Curve Clustering and Prediction. IEEE Transactions on Smart Grid 9, 5196–5206. https://doi.org/10.1109/TSG.2017.2683461

Acknowledgements

The authors gratefully acknowledge funding for this research from the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/M024652/1.