Applied Energy 260 (2020) 114246

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Cluster analysis and prediction of residential peak demand profiles using occupant activity data

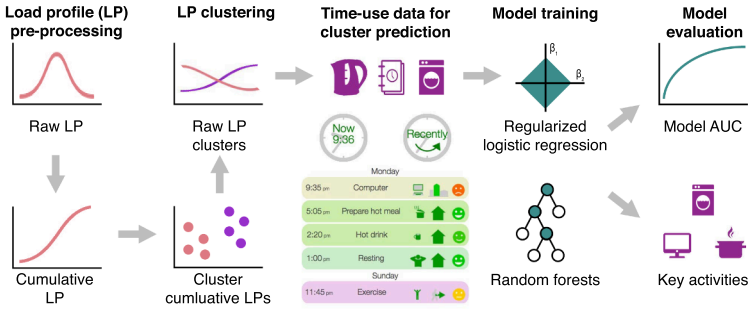
Aven Satre-Meloy*, Marina Diakonova, Philipp Grünewald

Environmental Change Institute, University of Oxford, South Parks Road, Oxford OX1 3QY, UK

HIGHLIGHTS

- Assembles data set of 269 hourly load profiles and occupant time-use data from UK households.
- Clusters cumulative load profiles to capture their full shape during evening peak hours.
- Predicts load profile cluster membership using occupant-reported activity data.
- Applies regularized logistic regression and random forests for load profile classification.
- Finds that cooking/eating, screen time, and laundry are key predictors of evening usage patterns.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Residential electricity demand
Cluster analysis
Regularization
Peak demand
Demand response
Time-use data

ABSTRACT

Researching the dynamics of residential electricity consumption at finely-resolved timescales is increasingly practical with the growing availability of high-resolution data and analytical methods to characterize them. One methodological approach that is popular for exploring consumption dynamics is load profile clustering. Despite an abundance of available algorithmic techniques, clustering load profiles is challenging because clustering methods do not always capture the temporal aspects of electricity consumption and because clusters are difficult to explain without additional descriptive household data. These challenges limit the use of cluster analysis to better understand behavioral and other drivers of electricity usage patterns.

We address these challenges by applying a novel clustering approach to a unique data set of high-resolution electricity and occupant time-use data from UK households. We cluster cumulative rather than raw load profiles to capture their full shape. Our clustering approach identifies two distinct patterns of electricity consumption during evening weekdays (5–9 p.m.), which are primarily differentiated by the timing of their peak demand. Next, we apply several classification algorithms to assess the potential for using time-use activity data to predict membership in these distinct usage clusters. The methods we use are suited to this predictive modeling context and are able to identify key activities driving patterns of electricity demand. We discuss how such an approach can inform more targeted strategies for residential peak demand reduction and response interventions as well as improve our understanding of constraints and opportunities for demand-side flexibility in the residential sector.

* Corresponding author.
E-mail address: aven.satremeloy@ouce.ox.ac.uk (A. Satre-Meloy).

https://doi.org/10.1016/j.apenergy.2019.114246
Received 6 September 2019; Received in revised form 20 November 2019; Accepted 23 November 2019

1. Introduction

The residential sector accounts for around 30% of total UK electricity consumption and 50% of UK national peak electricity demand [1,2]. 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 important in energy research, especially in the residential sector where temporal variability is high [3]. Reducing system peak demand lowers the costs and carbon-intensity of electricity generation [4]. 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 [5,6].

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 [7]. Access to high-resolution electricity consumption and activity data can strengthen these insights if appropriate methods are used to characterize 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 [8–13]. Cluster analysis is an algorithmic approach used to identify homogeneous groupings of data where no *a priori* grouping exists [14]. 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.

Cluster analysis of load profiles in the residential sector can be used for numerous practical purposes, including identifying characteristics that correlate with different energy usage patterns [11,15], creating more accurate profiles for groups of utility customers [16,17], which can then be used to design more appropriate tariffs [18], or targeting customers that may be particularly viable for demand response schemes [8,19,20]. These applications of cluster analysis can aid utility program designers in discovering potential energy saving opportunities for customers based on their specific patterns of usage while also enabling them to better coordinate demand-side resources for energy management.

Despite the emergence of this research, however, identifying representative classes of customers that can be explained by socio-demographic or other household data remains challenging for several reasons. First, clustering results have been shown to be highly sensitive to numerous methodological considerations and assumptions [21]. Especially in the context of clustering load profiles, comparatively few studies have addressed the issue that cluster analysis of time-series electricity data using conventional distances does not adequately capture temporal variations in consumption patterns [22–24].

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 for utilities or policymakers. In the context of targeting customers for demand response schemes, designing better tariffs, or identifying characteristics associated with energy usage behaviors, it is essential that clusters can be explained by associated descriptive data from the dwelling and its occupants.

Numerous studies have highlighted the need for additional information from households to verify the social, lifestyle, and behavioral patterns that underpin temporal variations in electricity consumption [8,10,20,25]. These studies stop short of investigating the social forces that influence the magnitude and timing of consumption because of data availability issues, and they instead make assumptions or simple guesses about the lifestyle and activity factors that explain different clusters [18,26,27]. Empirical, data-based evidence, rather than evidence from models based on assumptions or guesses as to which end-uses are driving consumption, is lacking.

This paper addresses these two challenges through an integrated cluster and classification analysis of evening peak-period electricity consumption in UK households. It introduces a novel data set of high-resolution electricity consumption and detailed time-use activity data and applies cluster and classification analyses to these for the purpose of identifying key activities that drive different patterns of electricity use from 5 to 9 p.m. in UK homes. It makes two primary contributions to the literature on household load profile segmentation and prediction.

First, it makes a methodological contribution by introducing a novel approach for pre-processing load profiles prior to clustering that focuses the cluster analysis on capturing temporal variations in electricity consumption. We propose clustering cumulative rather than raw load profiles, which enables more appropriate use of Euclidean distances to properly account for temporal differences between clusters. This approach to clustering load profiles does not appear to have been used in previous research. After we apply this pre-processing step, we cluster profiles using several different Euclidean distances and two types of clustering algorithms and then select the best performing methods, as evaluated using several cluster validity indicators (CVIs), to determine the optimal cluster assignment for our data.

The paper's second contribution is its incorporation of detailed time-use activity data from household occupants to predict cluster membership. We apply two classification models, which are specifically chosen for their ability to handle the complexities inherent to incorporating thousands of reported activities in the prediction task, as well as for their ability to identify key activities associated with distinct usage patterns. The aim of this stage of the analysis is to show how access to household time-use data can improve segmentation of customers and identify activities driving their demand. To the authors' knowledge, no previous studies have combined load profile segmentation via cluster analysis with time-use activity data in a predictive context to identify activities driving temporal variations in household electricity demand. This novel approach can inform efforts to develop more targeted and effective strategies for peak demand reduction and demand response in the residential sector.

The paper is organized as follows: Section 2 reviews related work, both on clustering electric load profiles and using time-use activity data in models of building electricity consumption. Section 3 describes the data used for subsequent analyses. Section 4 describes the clustering and classification analyses undertaken in this paper. Section 5 presents results, and Section 6 discusses implications of these and avenues for future research.

2. Related work

2.1. Cluster analysis of electric load profiles

There is a large and active literature on electric load profile segmentation using cluster analysis. In this review, we mostly focus on the empirical findings of papers that have used this approach, but we note that the literature reviewing various methodological approaches and considerations is equally active. The authors recommend [14,28,29] for reviews of this literature.

We summarize previous empirical findings in Table 1. As shown in the table, studies are concentrated primarily in Europe and the U.S., with one study each from China, South Korea, and South Africa. Numerous studies used the same data sets, such as the Irish Commission for Energy Regulation (CER) smart metering trial data. The residential sector is the most commonly analyzed. In terms of temporal resolution, most studies use data sampled at either 15 min, 30 min, or hourly, and data is in most cases sampled for one part of the year only (e.g., summer or fall). Sample sizes for these studies vary considerably. While cluster analysis in other disciplines is usually applied to large data sets, for load

Table 1
Summary of previous electric load profile clustering papers reviewed.

Study	Country	Building sector(s)	Data period and resolution	Sample size	Summary[a]		
					K	HD[b]	SM
Albert and Rajagopal [8]	U.S.	Residential	March-October 2010; 10-min	952	8	AF, SDF	x
Azaza and Wallin [31]	Ireland	Residential	2009–2010	4232	5		
Cao et al. [19]	Ireland	Residential	July 2009-December 2010; 30-min	4000	14		
Chicco et al. [32]	Romania	Commercial, Industrial	Date not given; 15-min	234	15		
Dent et al. [33]	UK	Residential	3 months in summer & winter; hourly	93	9		
Figueiredo et al. [30], Rodrigues et al. [34]	Portugal	Residential, Commercial, Industrial	3 months in summer & winter; 15-min	165	6–9		x
Flath et al. [18]	Germany	Residential, Commercial	Winter 2010–2011; 15-min	215	10–14		
Gouveia and Seixas [35]	Portugal	Residential	July-August 2014; 15-min	265	10	AF, DF, SDF	
Granell et al. [36]	UK & Bulgaria	Residential	2010; various temporal resolutions	197	2–4		
Granell et al. [9]	UK & Bulgaria (res), UK (com)	Residential, Commercial	April-November 2010 (res), 2009–2010 (com); 1-min (res), 30-min (com)	197 (res), 1207 (com)	4 (res), 6 (com)	DF, SDF	x
Haben et al. [26]	Ireland	Residential	2009–2010; 30-min	3622	10		
Iglesias and Kastner [22]	Spain	Commercial	August 2011- January 2012; hourly	5	n/a		
Jin et al. [37]	U.S.	Residential	June 2011-May 2012; hourly	100,000	15–30		
Kwac et al. [38,20,10]	U.S.	Residential	April 2008-October 2011; hourly	220,000	n/a		
Liu et al. [39]	Finland	Residential	2009; quasi-daily	11,964	4	DF, SDF	
McLoughlin et al. [11]	Ireland	Residential	July-December 2009; 30-min	3,941	10	AF, DF, SDF	x
Panapakidis et al. [40]	Greece	Commercial	January 2010-December 2011; 15-min	27	15		
Piao et al. [23]	South Korea	Residential, Commercial, Industrial	January 2007; hourly	1,205	4		
Räsänen et al. [16]	Finland	Residential, Commercial, Industrial	2008; hourly	3,989	19		
Rhodes et al. [12]	U.S.	Residential	November 2012-October 2013; hourly	103	2	AF, DF, SDF	x
Smith et al. [25]	U.S.	Residential	Summer 2011; 15-min	6,662	6		
Teeraratkul et al. [27]	U.S.	Residential	July-August 2012; hourly	1,057	12		x
du Toit et al. [24]	South Africa	Residential, Commercial Industrial	Summer 2007–2014; 30-min	Not given	2–3		
Tsekouras et al. [41]	Greece	Industrial	2003; 15-min	93	8–12		
Viegas et al. [13]	Ireland	Residential	2009–2010; 30-min	3,440	4	AF, DF, SDF	x
Xu et al. [42]	U.S.	Residential	January 2015; 15-min	Not given	9		
Zhang et al. [43]	China	Residential, Commercial, Industrial	August 2009; 30-min	155	5		

[a] Optimal number of clusters (K); Links household data (HD); Conducts statistical modeling (SM).
[b] Appliance factors (AF); Dwelling factors (DF); Socio-demographic factors (SDF).

profile segmentation there are numerous studies with samples in the hundreds of customers. Several, however, have much larger sample sizes in the hundreds of thousands.

The number of clusters (K) determined as optimal in different studies is highly variable and often dependent on the clustering algorithm used and nature of the data. Still, there does seem to be some consistency across studies with most finding 5–15 clusters. In numerous cases the authors do not select the number given as optimal by a CVI but instead use some judgment to determine a cut-off in number of clusters in light of the segmentation goal or to better match numbers that would be useful for electricity companies [e.g., 11,30]. Finally, the right-most columns in Table 1 indicate whether and what types of auxiliary household data (HD) are linked to clusters and whether or not these data are used in statistical models (SM) to predict cluster assignment. About a third of studies reviewed take this additional step after clustering.

From the papers reviewed, we generally find that cluster analysis for load profile segmentation is done for several different but often related purposes. These include, in order of frequency in the literature reviewed, (1) comparative analyses to determine which combinations of distances and clustering algorithms perform the best for load profiles given varying approaches for validating cluster results or different temporal resolutions for the data [19,22–24,27,29,32,34,36,37,41]; (2) studies of stability of cluster assignments or entropy of load shapes for individual households, which measures how often households consume the same daily load shape pattern [8,20,23–25,27,38,42,43]; (3) analyses that use statistical methods to link load profile clusters to dwelling or other household data [9,11–13,30,35]; (4) investigations of the variability in timing of peak demand, the contributions of different customer segments to peak demand, or related time-of-day and seasonal effects on electricity consumption patterns [25,31,39,42]. Given that this paper's aims are primarily related to (3) and (4), several findings specific to these aims from the literature are given here. Studies that have used statistical models to associate household data with cluster assignment show that several dwelling, socio-demographic, and appliance factors are influential in predicting load profile cluster. Granell et al. [36] find that number of bedrooms and occupants are key drivers, while Gouveia and Seixas [35] find that dwelling year of construction, floor area, and heating and cooling equipment, along with number of occupants and monthly income, are important. In Viegas et al. [13], the authors find customer employment status and age are significant, as well as number of dishwashers, electric cookers, and washing machines. McLoughlin et al. [11] associate dwelling, socio-demographic, and appliance ownership variables to 10 profile clusters, finding that factors such as number of bedrooms, ownership of energy-intensive appliances, and age of head-of-household (HoH) are significant. The authors note, however, that in many cases, standard errors for regression coefficients are large owing to small samples in certain variable sub-categories. Finally, Rhodes et al. [12] find that variables such as working from home, hours of television watched per week, and education levels have significant relationships with average load profile shape but that these are variable across seasons.

2.2. Prediction and classification of building electricity consumption

Cluster analysis is often used as a pre-processing step for predicting or classifying building electricity consumption, and there exist numerous data-driven approaches that can be applied to the clustering results in a predictive context. Given that we combine both cluster analysis and classification in an integrated analysis in this paper, in this section we provide a brief review of the literature on building energy prediction and classification.

Wei et al. [44] present a recent and detailed review of the main categories of approaches for building energy prediction. These categories are described as white-box, grey-box, and black-box approaches. White-box approaches refer to physics-based models that use detailed building characteristics, such as appliance power ratings, and thermodynamic principles to simulate building operations and predict energy consumption. These approaches do not require historical energy consumption information and have high accuracy but also require detailed data measurements of building characteristics, which makes them more computationally intensive [45]. Grey-box approaches require both physical building information as well as historical data and combine statistical methods with building physics models to predict building electricity consumption [44]. Conditional demand analysis (CDA) is a specific method that performs regression to determine the use level of individual appliances [46]. While these methods benefit from the use of statistical methods and thus can be applied to larger samples, one disadvantage is that they involve complex interactions between end-uses, which increases the uncertainty of their predictions. The final category, black-box approaches, refers to methods that predict building electricity consumption strictly using historical data and statistical analysis. These approaches are becoming increasingly popular for their ability to incorporate a wide range of data types, including occupant behavior and socio-demographic factors, without sacrificing predictive accuracy [47].

As this paper's classification approach falls under the black box or "data-driven" family of building energy consumption prediction, a short review of several common methods within this family is provided here. Some of the more well-known data-driven methods include various types of regression (e.g., multiple linear regression, polynomial regression) as well as more complex non-linear methods such as support vector machines (SVM), artificial neural networks (ANN), and decision trees (DT). In a systematic review of the data-driven building energy consumption literature, Amasyali and El-Gohary [48] found that around half of the studies reviewed employ ANNs, a quarter use either SVM or regression methods, and only around 4% use decision trees. Due to their highly non-linear nature, both ANNs and SVMs and their extensions are able to achieve high accuracy in energy consumption prediction tasks [e.g., 48]. Regression approaches have historically showed lower predictive accuracy than ANNs and SVMs because of their inflexibility [44], but recent extensions to standard multiple linear regression, such as regularization methods, deliver performance gains in terms of prediction and model interpretability [7,21]. DTs have been shown to perform as well as ANNs and SVMs in some contexts [50,51], though in general their predictive accuracy is somewhat lower [44].

There are many other data-driven algorithms for building energy consumption, and a more detailed review is beyond the scope of this paper. We recommend both Wei et al. [44] and Amasyali and El-Gohary [48] for more in-depth reviews of these approaches. As will be described in Section 4, we apply two of the four types of data-driven methods reviewed above—regression and decision trees. Our approach, however, goes beyond previous research by applying extensions to these two methods, which are necessary given the complexities inherent in our predictive modeling context.

2.3. Time-use activity data in electricity demand models

This paper uses time-use data to predict load profile cluster membership, so to familiarize readers with this area of research, in this section we provide a brief review of time-use data and its application to energy demand modeling. For a more complete review, we direct interested readers to Torriti [52].

Time-use activity data have been collected as part of national studies on how people spend their time [53,54]. More recently, these data have been used as inputs to energy demand models [55–59]. Studies investigating the associations between time-use and electricity use have found some consistency in terms of the activities that influence electricity use patterns. Key activities linked to higher consumption include meal preparation-related activities and meal times [60–63], housework and personal time [58,64], and some recreational activities, especially TV use [7]. Activities that are linked with lower consumption unsurprisingly include sleeping and resting [61,62].

Approaches for incorporating time-use data into energy demand models vary widely, and a detailed review is beyond the scope of this paper. Key limitations, however, have been summarized in previous literature [52,65]. Among these, several are particularly relevant for this paper. First, as McKenna et al. [65] note, time-use studies were not originally designed for energy demand modeling, so time-use data often are not tailored to questions regarding electricity use. For instance, study participants are not asked to differentiate between energy-intensive or low-energy alternatives of the same activity (e.g., cold meals versus hot meals). Second, as time-use data have historically been collected using paper diaries that ask participants to write a primary and secondary activity during each 10-min window of the day (the 'time budget' approach), overlapping activities were not easily accounted for, and some electricity-relevant activities that may occur during shorter time periods but might still be important for understanding demand (e.g., boiling the kettle) were underreported. Furthermore, as Torriti [52] notes, another key challenge is using activity data to model multiple-person households [cf. 65].

The data collection process described in the following section was designed with these limitations in mind and addresses these through a new approach to collecting time-use data.

3. Data

The collection process for the data analyzed in this paper has been described in previous work [7,66]. We briefly review the study design and data collection methods and then describe several updates to the data and specific sampling considerations for this paper's analysis.

Data are collected as part of a five-year study for which data collection is ongoing [67]. 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. Any household member above the age of 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 h, from 5 p.m. until 9 p.m. on the following day, thus covering two typical peak demand periods in the early evening. The UK is winter-peaking, and around two-thirds of data are therefore collected between September and April.

Activities are recorded using a purpose-built app that comes preinstalled 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. A schema representing an example activity entry sequence is shown in Fig. 1. The 'Activity' screens are tailored to collect numerous details on the type of activity, appliances used, and other energy-relevant details.

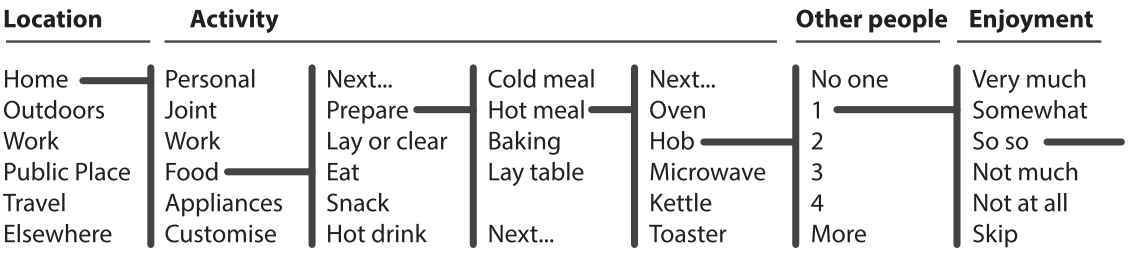


Fig. 1. Example of an activity entrance sequence on the activity recorder [70].

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 [68]. 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. These capabilities facilitate reporting of overlapping activities as well as activities that may not fit into the typical 10-min reporting windows used in past time-use studies. Previous analyses have examined the accuracy of activity reporting in this data set and have found around 80% of reports for certain activities are entered within 10 min of the activity itself [see 65]). A description of the data storage and handling procedures is given in Grünewald and Diakonova [69].

In addition to the procedures mentioned above, we exclude households with the following from the sample:

- poor electricity readings (e.g., failure to attach the recorder);
- missing electricity readings between 5 and 9 p.m. Multiple imputation is unnecessary due to the low incidence of missing data ($N = 14$, ~5% of sample);
- no activities reported between 5 and 9 p.m.;
- PV (unreliable net-demand electricity readings);
- weekend participation dates (weekend demand tends to be lower and different to weekday demand [30]).

4. Methods

In this section, we describe our methods for clustering 5 to 9 p.m. load profiles and then predicting load profile cluster using occupant activity data. The overall analysis approach taken in this paper is most similar to the approaches demonstrated in [11–13,30]. The key difference is our use of occupant activity data for the predictive modeling task and our application of regularization methods for identifying key predictive activities. A graphical overview of our methodological approach is shown in Fig. 2 and explained step-by-step in Sections 4.1–4.3. Section 4.1 describes how we approach the clustering task and then pre-process and normalize the data, including the assumptions we make at these steps. Section 4.2 explains our procedure for validating cluster results and assigning cluster membership. Section 4.3 then describes the predictive modeling component of our analysis, which uses occupant activity data to predict cluster membership. In this section we also describe our model training and evaluation procedures.

4.1. Data pre-processing and normalization

In this paper we take a raw-data-based approach [11,12,16,27,42] and apply clustering algorithms to whole time-series electricity data rather than extracted features [e.g., 10,31,26]. We downsample our raw electricity readings to create hourly profiles[1] where the hourly average is assigned to the interval beginning hour (e.g., electricity use data from 5:00 to 5:59 p.m. are averaged and assigned to 5:00 p.m.). We then extract the four values corresponding to the hourly averages from 5 to 8 p.m. for day 1 and day 2 of the study period so that each household can be represented by two 4-element vectors, given by:

$$l_i = \begin{pmatrix} P_{d5} \\ P_{d6} \\ P_{d7} \\ P_{d8} \end{pmatrix} \quad (1)$$

where l_i is the load profile for each home i, and P_{dh} is the hourly power consumption in Watts (W) of home i on day d at hour h.

The two load vectors per household (day 1 and day 2) are treated as independent in this analysis in order to increase the sample size. Fig. A.1 in the appendix shows that when only day 1 profiles are analyzed, results for cluster membership and shape are similar to those obtained on the combined sample. This step improves the performance of both the cluster and classification algorithms we use, as these are typically better suited to larger sample sizes.

Our final sample size consists of $N = 269$ household load profiles. The normalization technique we use is that proposed by Jin et al. [37] to normalize profiles by first subtracting minimum demand from each hourly value ('de-minning') and then dividing each hour's consumption by the 'de-minned' total. For our 5–9 p.m. profiles, we subtract the minimum over this four-hour period rather than the daily minimum. After normalization, each household load profile represents hourly discretionary usage. We use this normalization technique because our focus is segmenting households by the temporal variation in their electricity use in order to identify different shapes of peak-period usage profiles. We expect de-minning is especially appropriate for linking occupant activities to electricity consumption patterns, as activities are most likely drivers of discretionary changes in usage (as opposed to baseload demand). De-minning also mitigates distortion of load profiles for households with high baseload [71].

We take one final data-preprocessing step to more appropriately capture the temporal variations in de-minned hourly load profiles during clustering. When clustering time-series data, it is important to account for their ordered, temporal nature. Even though distances in Euclidean space are used widely for clustering time-series data, these do not capture temporal variations and thus are inadequate for determining dissimilarity of time-series vectors. A simple example of this issue is given in Fig. 3, which plots three 6-element vectors with magnitude 5 as raw time series. Using a Euclidean distance to calculate the pairwise distance for each pair of these vectors results in the same distance equal to $500\sqrt{2}$. In other words, when using a Euclidean distance, these vectors are equidistant from one another. However, given their overall shape, we would likely say profiles a and b are more similar to one another than they are to profile c.

Several studies have noted the issue of Euclidean distances not appropriately handling time-series load data [24,27]. We propose a novel solution to this problem: instead of clustering the raw time-series data, we take the integral of the load shape l_i, constructing a new sequence C_i of the same length that gives the cumulative load, where each component is of the form:

[1] We use hourly averages because we do not find clear performance gains in the clustering task when sampling our data at higher resolutions, such as 30-min or 15-min [cf. 35], and because using hourly averages reduces the dimensionality of the data considerably.

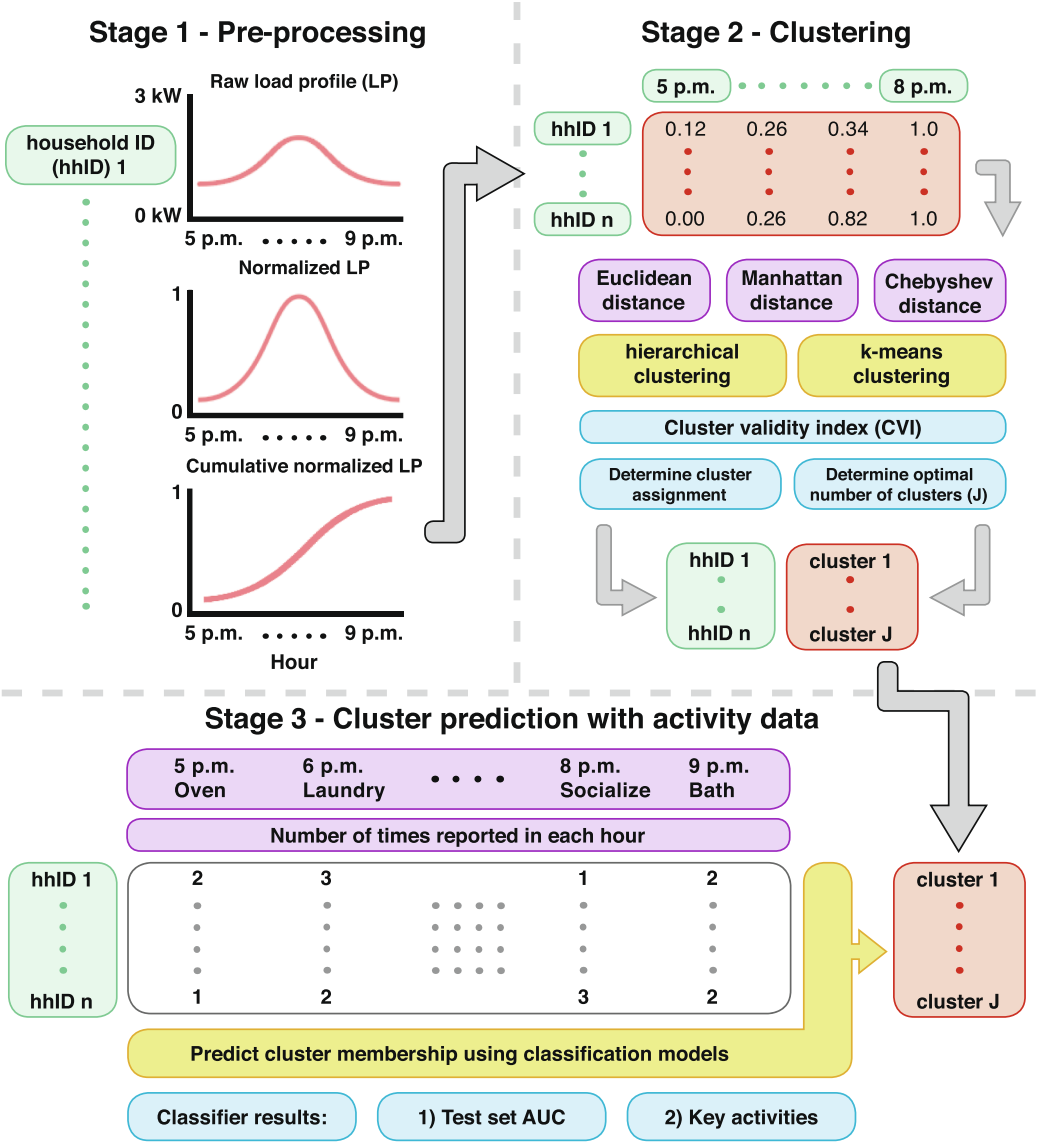


Fig. 2. Overview of methodological approach for clustering and classifying load profiles, with Stages 1, 2, and 3 described in Sections 4.1–4.3, respectively.

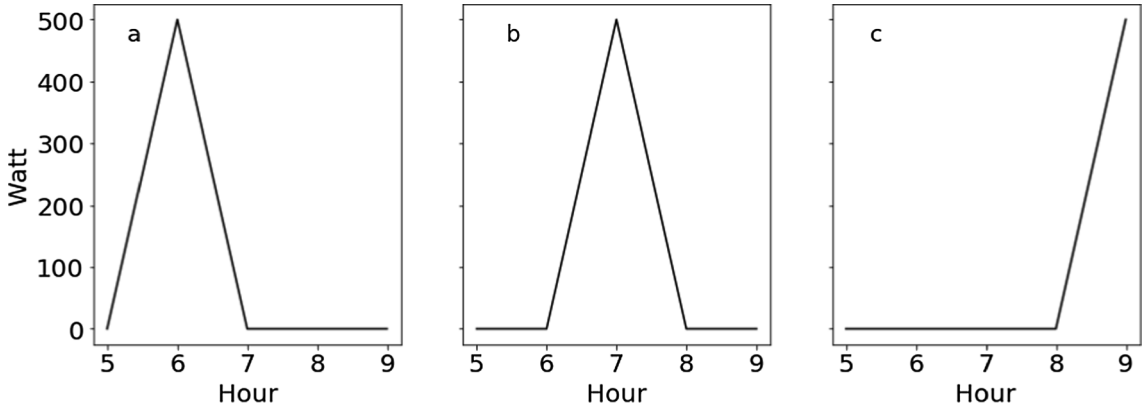


Fig. 3. Example of three hypothetical load profiles with equal Euclidean distances between each pair.

$$C_i = \sum_{h=5}^{8} P_h \tag{2}$$

where C_i is the cumulative load profile for each home i, and P_h is the hourly electricity consumption (W) of home i at hour h. Returning to the example above, the new cumulative load profiles shown in Fig. 4 are no longer equidistant. The Euclidean distance between d and e is 500, the distance between d and f is 1000, and the distance between e and f is $500\sqrt{3}$. Clustering the cumulative load profiles would thus group together d and e, which better accounts for the similarities in

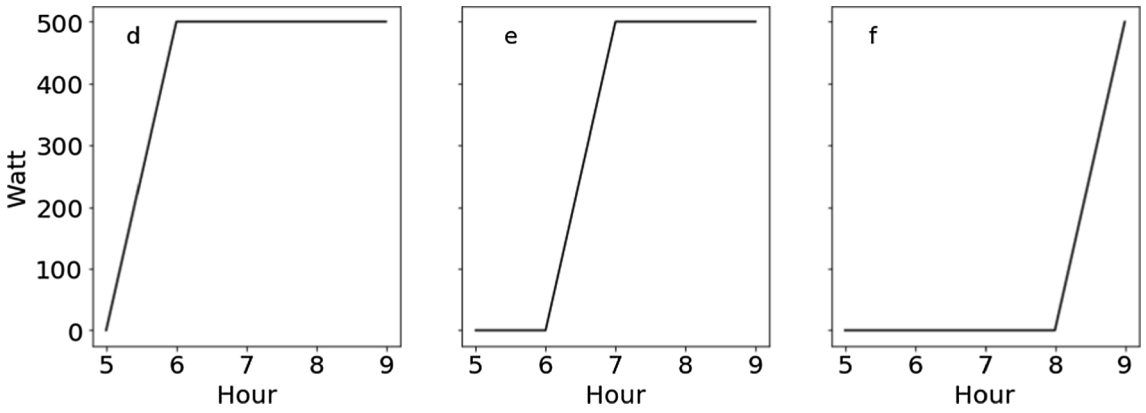


Fig. 4. Cumulative load profiles with Euclidean distances such that d and e are closer together than either is to f.

Table 2
Algorithms used to cluster peak-period load profiles.

Clustering algorithm	Equation	Description
k-means	$\underset{S}{\operatorname{argmin}} \sum_{k=1}^{J} \sum_{x \in S_k} d(x, \mu_k)^2$	This method minimizes the distance between each household load profile x and the groups μ_k.
Single (hac)	$\min_{ij} d(X_i, Y_j)$	This method minimizes the distance between the closest members of the two clusters.
Complete (hac)	$\max_{ij} d(X_i, Y_j)$	This method minimizes the distance between the most dissimilar or furthest apart members of the two clusters.
Average (hac)	$\frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(X_i, Y_j)$	This method minimizes the average distance between all pairs of members of the two clusters.
Centroid (hac)	$\|\|c_A - c_B\|\|$	This method finds the mean vector location for each of the clusters and takes the distance between these two centroids.
Ward (hac)	$\Delta(A, B) = \sum_{i \in A \cup B} \|\|\vec{x_i} - \vec{m}_{A \cup B}\|\|^2 - \sum_{i \in A} \|\|\vec{x_i} - \vec{m}_A\|\|^2 - \sum_{i \in B} \|\|\vec{x_i} - \vec{m}_B\|\|^2 = \frac{n_A n_B}{n_A + n_B} \|\|\vec{m}_A - \vec{m}_B\|\|^2$ This method minimizes the total within-cluster sum-of-squares. Clusters are combined such that their merger results in minimum information loss.	

where:

•for *k*-means, the centroid of cluster S_k is its mean vector $\mu_k = \frac{1}{|S_k|} \sum_{x \in S_k} x$,

• for single, complete, and average, $(X_1, \ldots, X_k)$ and $(Y_1, \ldots, Y_j)$ are the observations for clusters X and Y,

• for centroid, c_A and c_B are the centroids of clusters A and B, respectively,

• for Ward, $\vec{m}_j$ is the center of cluster j, and n_j the number of points in it. Δ is the merging cost of combining A and B.

their temporal shape.

While several other approaches have been proposed for clustering time-series data, such as the specific distance measures proposed in Liao [14], we argue our approach is advantageous both in its conceptual simplicity as well as its ease of implementation. Based on our review of existing load profile cluster analyses, this step is wholly absent from past work. Xu et al. [42] suggest a similar approach to integrate load shapes, but they do so to consider maximum demand over the time series rather than to capture the temporal shape of load profiles.

After these pre-processing and normalization steps, we next move to clustering these de-minned, cumulative load profiles where we are able to achieve the simultaneous goals of focusing on temporal variations in discretionary usage while also enabling straightforward use of Euclidean distances.

4.2. Cluster analysis of peak-period load profiles

In this section we describe our approach to cluster peak-period, de-minned, cumulative load profiles. Because numerous studies have identified the issue of unstable clustering results when using just one clustering algorithm to determine the optimal number of clusters [21,29], we address this issue by using the *NbClust* package in R to compare clustering results across numerous distance and algorithm combinations. Specifically, we compare three measures of Euclidean distance, two types of clustering algorithm, and six cluster validity indices (CVIs). Our aim is to ensure the cluster assignment is robust to numerous methodological considerations [cf. 22,41].

The distance measures used are three variants of the Minkowski metric, which is defined as:

$$D(X, Y) = \sqrt[p]{\sum_{i=1}^{n} |(x_i - y_i)^p|} \tag{3}$$

where $D(X, Y)$ gives the distance of order p between two points $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$. The simple Euclidean distance is given when $p = 2$, the Manhattan or city block distance when $p = 1$, and the Chebyshev or maximum distance when $p = \infty$. In addition to our motivation to use Euclidean distances given the pre-processing step taken to cluster cumulative load profiles, these distance measures are some of the most frequently used in previous research and have shown improved clustering results in terms of stability and performance across various CVIs [22,24,32,37].

We include two types of clustering algorithms in our comparative analysis. These are *k*-means and several variants of hierarchical agglomerative clustering (hac). From the studies reviewed, these were the

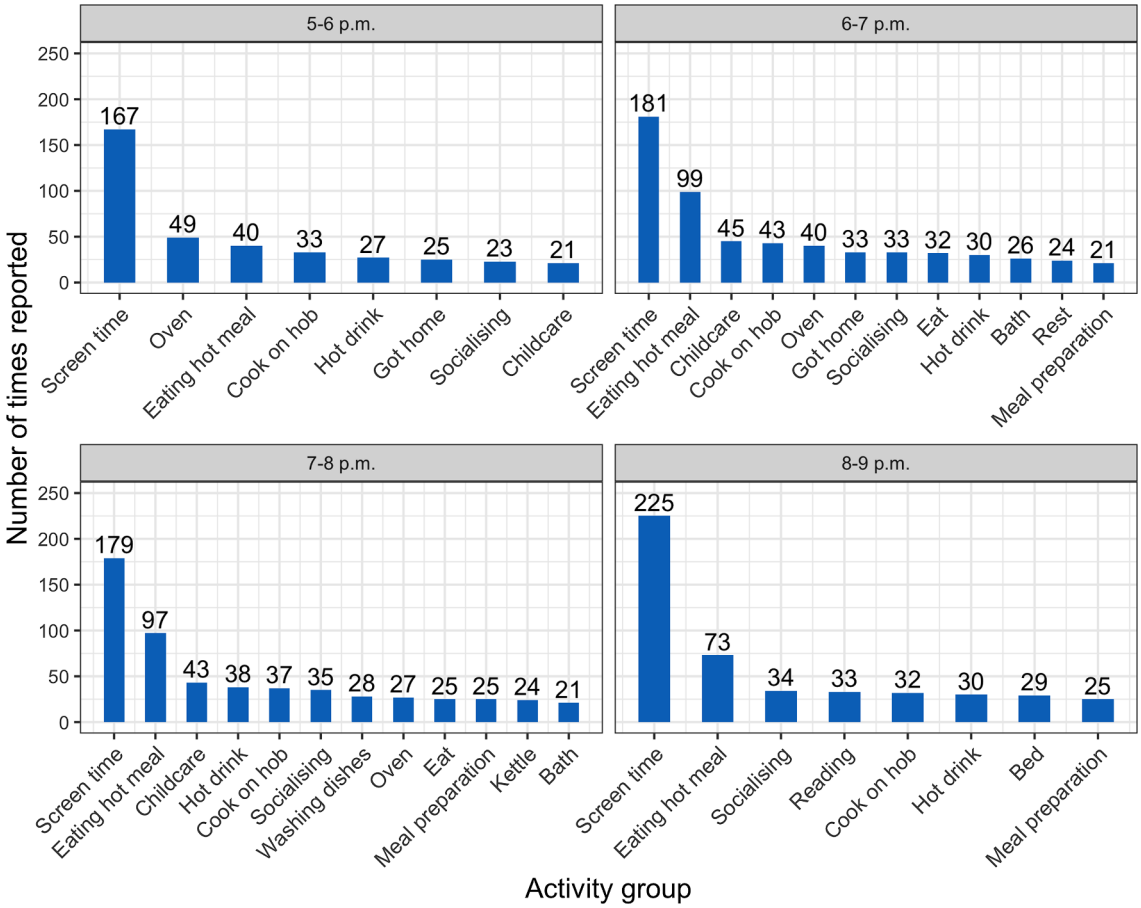


Fig. 5. Frequency histogram of reported activity groups by hour of evening peak period. Activities reported fewer than 20 times are excluded for brevity.

most commonly applied, and numerous studies also concluded that these perform well for load profile clustering [9,32,34,37,41]. For the hierarchical clustering, we use several linkage criteria, which determine how clusters are merged together at successive steps. A list of the algorithms used, along with their formulas and descriptions, is given in Table 2.

For each combination of distance measure and clustering algorithm, we evaluate the clustering results using several CVIs to determine the optimal number of clusters and cluster assignment. There are many different CVIs reported in the clustering literature, and for load profile clustering specifically, there are several that are used quite frequently. A list of the CVIs we validate our clustering results with is given in the appendix (Table A.1).

To determine the optimal number of clusters, we again use the *NbClust* package, which returns the optimal number as selected by each of the combinations of distance and algorithm. We constrain the number of clusters to between 2 and 20 and do not evaluate results for larger numbers of clusters given our sample size and segmentation objective. While a comparative analysis of clustering results across the different indices (in terms of cluster assignment and optimal number of clusters) is interesting from a methodological perspective, doing so is beyond the scope of this paper.

Instead, to determine the optimal number of clusters and then assign cluster membership, we select the CVI that shows the least variability in index scores across the combinations of distances and algorithms and for varying numbers of clusters. We take this approach because some indices show widely varying results in index scores across different methods. A less variable index suggests clustering results are robust to different algorithms and distances. To assign final cluster membership, we use the most common clustering assignment across the different distances and algorithms, given the selected CVI and the number of clusters shown to be optimal for the data.

4.3. Classification of cluster membership using occupant activity data

The final stage of our analysis aims to train models to predict the cluster membership of a new household, as determined by its peak-period electric load profile, given the activities reported by its occupants. For this analysis, we train two different classifiers. The first is a linear binary logistic regression classifier with elastic net regularization, and the second is a non-linear random forests classifier. Each of these is described in Sections 4.3.2 and 4.3.3. Before we introduce these methods, however, we first explain how a generic classification algorithm works and how we prepare our activity data for model training.

4.3.1. Classification model setup

A classifier is a function f that maps a set of predictors p of an observation, in this case a set of activities associated with a given load profile, to a categorical response variable y, which represents one of the J load profile clusters. It is given by Eq. (4).

$$f: \mathfrak{R}^p \mapsto y \tag{4}$$

$$y \in \{c_1, c_2, \ldots, c_j\}.$$

Previous studies that have trained classifiers to predict load profile cluster membership have used various predictors or features, such as household socio-demographic data, physical characteristics of the dwelling, appliance ownership, other occupant-related survey data, and extracted features from load profiles [8,9,11–13,27]. To the author's knowledge, this is the first paper that attempts to predict load profile cluster membership using occupant time-use data.

We set up our training data in the following way: first, we group individual time-use codes (TUCs) into categories of activities that are similar. For instance, while there are separate TUCs for 'Screen time' to indicate whether a TV, tablet, computer, or other device is used, we group these into one category. Doing so has the benefit of reducing 113 unique TUCs into 61 activity groupings. In total, $N = 3038$ activities are reported in our sample of households during the hours of 5 to 9 p.m. (across both the first and second days of each household's study participation).

Next, for each household and each hour from 5 to 9 p.m., we count the number of times an activity is reported. In doing so, we create an 'hour-activity' predictor variable for each household to indicate the number of times a given activity group is reported each hour. We aggregate activities at the hour level to prevent the predictor matrix from getting too large. Even still, given the number of hours and activity groupings, the matrix $\mathfrak{R}^p$ has dimensions 269×177, where the number of weekday load profiles is $N = 269$ and the number of 'hour-activity' predictors is $N = 177$. Fig. 5 gives activity counts for all activities by the hour in which they are reported.

This approach to the predictive modeling task presents some specific analytical challenges. First, the predictor matrix is relatively wide, and although it is not 'high-dimensional' in the sense of $p \gg n$, the inclusion of such a large number of predictors in a conventional linear model creates challenges for interpretability as well as makes the model more prone to overfitting [72]. Second, because the predictors are activity counts and because many activities are not commonly reported, the predictor matrix is also very sparse with many zero values. This issue raises to two additional challenges, the first of which is that in a two-class modeling scenario with wide data, the classes will almost always be linearly separable, meaning the outcome variable separates a predictor or combination of predictors completely or quasi-completely [73]. When this is the case, typical linear modeling approaches, such as logistic regression, cannot be used. The second related challenge is that of multicollinearity, which occurs when two or more predictors are highly correlated [74]. Multicollinearity can cause issues of stability, where simple changes to the model or data lead to large changes in model parameters. As was the case with the former issue, the predictor matrix as it is designed in this analysis is prone to issues of multicollinearity given some activities are rarely reported and thus several hour-activity variables have near zero variance.

The specific classification methods described in the following sections are designed to address these challenges. We choose a linear method and a non-linear method to compare their performance in correctly classifying load profiles as belonging to clusters given activities reported by household occupants. Our objective in this part of the analysis is both examining the predictive power of activities while also identifying key activities that explain varying patterns of peak-period electricity consumption.

4.3.2. Logistic regression with elastic net regularization

Logistic regression is a Generalized Linear Model (GLM), a generalization of linear regression that allows for responses that are not continuous, quantitative variables. In the classification setting, especially when the dependent variable is binary, logistic regression is commonly used, and it has been applied to load profile classification in several previous studies [11–13]. We use logistic regression to determine the likelihood a load profile will be classified as belonging to a specific cluster given the activities occupants report each hour from 5 to 9 p.m.

In the case of a binary response coded as $Y \in \{0, 1\}$, linear logistic regression models the log-likelihood ratio as a linear combination of predictor variables, as in

$$\log\frac{\Pr(Y = 1|X = x)}{\Pr(Y = 0|X = x)} = \beta_0 + \beta^T x \tag{5}$$

where $X = X_1, X_2, \ldots, X_p$ is a vector of predictors, $\beta_0 \in \mathbb{R}$ is an intercept term, and $\beta \in \mathbb{R}^p$ is a vector of parameters or regression coefficients. The inverse of this transformation gives the conditional probability

$$\Pr(Y = 1|X = x) = \frac{e^{\beta_0+\beta^T x}}{1 + e^{\beta_0+\beta^T x}} \tag{6}$$

which, without restricting the parameters (β_0, β), necessarily constrains the probabilities to lie in [0, 1]. The logit transformation (Eq. (5)) transforms this conditional probability to a log-linear scale, which enables fitting the model parameters without constraints.

Fitting a logistic regression model is typically done by maximizing the likelihood, which is equivalent to minimizing the negative log-likelihood

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{-\frac{1}{N}\mathscr{L}(\beta_0, \beta; \mathbf{y}, \mathbf{X})\right\} \tag{7}$$

where $\mathbf{y}$ is the N-vector of outcomes (in a binary classification problem, $\mathbf{y} \in \{0, 1\}$), $\mathbf{X}$ is the $N \times p$ predictor matrix, and $\mathscr{L}$ is the log-likelihood function, as given in Eq. (5). The negative log-likelihood for logistic regression thus takes the form

$$\begin{aligned} &-\frac{1}{N}\sum_{i=1}^{N}\{y_i \log\Pr(Y = 1|x_i) + (1 - y_i)\log\Pr(Y = 0|x_i)\}, \\ &= -\frac{1}{N}\sum_{i=1}^{N}\{y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0+\beta^T x_i})\}. \end{aligned} \tag{8}$$

The model parameters or coefficients (β) in binary logistic regression are interpreted as the expected change in the $\log(odds)$ of $Y = 1$ for each one-unit increase in the corresponding predictor variable, holding other predictor variables in the model constant. To ease interpretation of coefficients, we convert the $\log(odds)$ to odds by finding the exponent $\mathrm{Exp}[\beta]$, and we can then find the increase or decrease in the likelihood of being in a particular load profile cluster given an increase in the number of activities reported in each hour of the evening peak period. Model coefficients for the logistic regression are reported both as $\log(odds)$ (β) and odds ($\mathrm{Exp}[\beta]$) for this reason. While we have not discussed specific details here, logistic regression can be extended to multi-class classification problems, as well.

Returning to the challenges inherent to our analysis approach, we address these through the application of regularization to the logistic regression model. Regularization, also called penalized regression, refers to a family of methods that regularize or constrain the parameter estimation process by adding a penalty term that shrinks the magnitude of model coefficients toward zero. There are several benefits to penalizing the estimation process in this way.

The first benefit is that doing so can significantly reduce the variance of regression coefficient estimates. Variance refers to the amount by which model coefficients change if they are estimated with different observational data. Ideally, these estimates should not change much between different data sets, but if a method has high variance, then small variations in the training data can result in large changes in model estimates. This is called overfitting, which often occurs when a very complex model is fit too specifically to the data it is trained on and thus does not generalize well to new data. Especially when p is large, fitting the full model typically increases both its complexity and the variance of its predictions.

In addition to the benefit of reducing the variance of estimates, a certain class of regularization methods have the benefit of sparsity, meaning they can estimate some of the coefficients to be exactly zero. In this way, these sparse statistical modeling methods can also perform variable selection. Sparsity in statistical models improves their interpretability and is especially useful in the case of large predictor sets. For more detail on statistical learning with sparsity, we direct interested readers to Hastie et al. [75], and for an overview of regularization methods and the benefits of applying these in statistical models of energy consumption, see Satre-Meloy [7].

While there are numerous types of sparse regularization methods, in this analysis we use the elastic net penalty, introduced by Zou and Hastie [76]. The elastic net penalty is a combination of two other regularization methods, the lasso and ridge penalties. Its lasso-related properties are the ability to set coefficients exactly to zero, thus yielding sparse models, but it also has properties that make it suitable both for handling data where classes are linearly separable and where there exist highly correlated variables [75]. The elastic net penalty is given by

$$\lambda \sum_{j=1}^{p} \{(1-\alpha)\beta_j^2 + \alpha\,|\beta_j|\} \tag{9}$$

where λ is a tuning parameter that determines the magnitude of the penalty that is applied and α is tuned to combine lasso and ridge penalties. If $\alpha = 0$, the coefficients are penalized by the sum of their squares (ridge or ℓ_2 penalty), and if $\alpha = 1$ they are penalized by the sum of their absolute value (lasso or ℓ_1 penalty). Setting α between these gives the elastic net penalty. When applied in the case of logistic regression, the general form of the penalized optimization problem for minimizing the negative log-likelihood is

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{-\frac{1}{N}\mathscr{L}(\beta_0, \beta; \mathbf{y}, \mathbf{X}) + \lambda \sum_{j=1}^{p} \{(1-\alpha)\beta_j^2 + \alpha\,|\beta_j|\}\right\} \tag{10}$$

where the notation is the same as in Eqs. (7) and (9) and the likelihood function $\mathscr{L}$ is as in Eq. (8). The elastic-net problem is convex and can be solved with numerous different algorithms. The R package *glmnet* uses coordinate descent and specifically a proximal-Newton iterative approach, which uses a quadratic function to repeatedly approximate the negative log-likelihood [77]. The predictors and response are centered and standardized before the algorithm is run.

Because both α and λ can be tuned across a range of values to find the best combination and magnitude of ℓ_1 and ℓ_2 penalties that minimizes the training error, cross-validation is typically used to find these. More details on this procedure are given in Section 4.3.4.

One final note on the use of regularization in logistic regression concerns the question of inference. We do not present typical inferential concepts such as confidence intervals and significance levels for model coefficients because these are not generally appropriate for use with regularization methods [78]. There is, however, a growing literature on post-selection inference [75,79,80], which aims to develop methods that can construct confidence intervals and significance tests for regularization methods. We do not take the additional step of significance testing in this paper given the non-random nature of our study sample.

4.3.3. Random forests

Tree-based methods for classification involve splitting the predictor space into regions that can be used to make predictions for new observations. These methods are so-called because they incorporate a set of rules used to partition the predictor space that can be summarized in a tree. Tree-based methods have been used in many instances for classification and regression related to load profile prediction [13,30,81,82].

Decision trees are a popular statistical learning method because of their simplicity and intuitiveness, but they are often prone to overfitting, which makes them less accurate for predictions. An ensemble method called random forests, which involves producing many decision trees from bootstrap aggregated samples of the data (called bagging) and then giving a consensus prediction across these, can greatly reduce the variance of predictions. Random forests is also able to handle high-dimensional data and is robust to outliers and imbalanced classes. The improved predictive performance using random forests over a single decision tree, however, comes at the expense of some interpretability, as it can be difficult to determine which predictors are most important in the model.

In the classification setting, a decision tree predicts each observation to belong to the most commonly occurring class of observations in a given region or node. Because most predictors cannot perfectly separate or predict the correct class for a given observation, they are considered 'impure'. To determine which separation is best given different predictors, a measure of impurity is used. While there are numerous ways to compare impurities, a popular measure is the Gini index [83], which is a measure of variance across K classes and is given by

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}) \tag{11}$$

$$\hat{p}_{mk} = \frac{1}{N_m}\sum_{x_i \in R_m} I(y_i = k)$$

where m is a node representing region R_m with N_m observations, and observations in node m are classified to class $k(m) = \text{argmax}_k \hat{p}_{mk}$, the majority class in node m. The Gini index takes on a small value when a node contains mostly observations from one class. At each split, the predictor with the lowest impurity is used to partition the predictor space until impurity does not decrease further with any subsequent split.

Because random forests works by resampling B bootstrapped training sets with replacement from the training data and then constructing B decision trees for each of these training sets, it is possible that the decision trees could be highly correlated, especially if there is a particularly strong predictor that is selected as the first (or root) node. For this reason, at each split in the tree a random subset of predictors rather than the full set is considered. The size of the random subset M of predictors p that can be selected at each split in the tree is typically chosen such that $M \approx \sqrt{p}$, but this is also a hyperparameter that can be tuned using cross-validation.

Finally, as mentioned above, interpretability is sometimes challenging when using random forests given the bagging procedure. However, we can still obtain an overall summary of the importance of each predictor using the Gini index by summing the total amount the Gini index decreases when split by a given predictor, averaged across the total number of B trees. The variable importance value for each predictor can then be normalized to a range of $[0, 1]$ for comparison.

4.3.4. Model training and evaluation procedure

Models are trained on a sample of 70% of the observations, which we subsequently refer to as the 'training set'. We hold out 30% of our data as the 'test set'. These splits are chosen to create a slightly larger test set than would be the case if we used a standard 80/20 split. As both classifiers require tuning of hyperparameters, we use k-fold cross-validation to tune and train the models. K-fold cross-validation involves splitting the training data into k subsets of equal size. Then each of the k subsets is left out and the other $k-1$ subsets are used to train the model. The left out subset is what the trained model predicts on, and some type of performance metric is computed. The k estimates of performance are averaged to get an overall training error. We use repeated k-fold cross-validation, which repeats the procedure above more than once, randomly resampling the k subsets or folds each time. This method often leads to a reduction in variance for test set predictions, and it is feasible with the relatively small number of observations in our training data. We use 5-fold cross-validation with 5 repeats.

The performance metric we use for all model training and testing is the receiver operating characteristic (ROC) and associated area under the ROC curve (AUC). While model accuracy is a straightforward method for describing classifier performance, it is a weak measure when the classes are imbalanced because it treats all classes equally, meaning if 80% of the training data observations belong to one of two classes, then simply predicting the dominant class for all observations would give an accuracy of 80%. Accuracy also does not assess whether the classifier's errors are more often false positives or false negatives

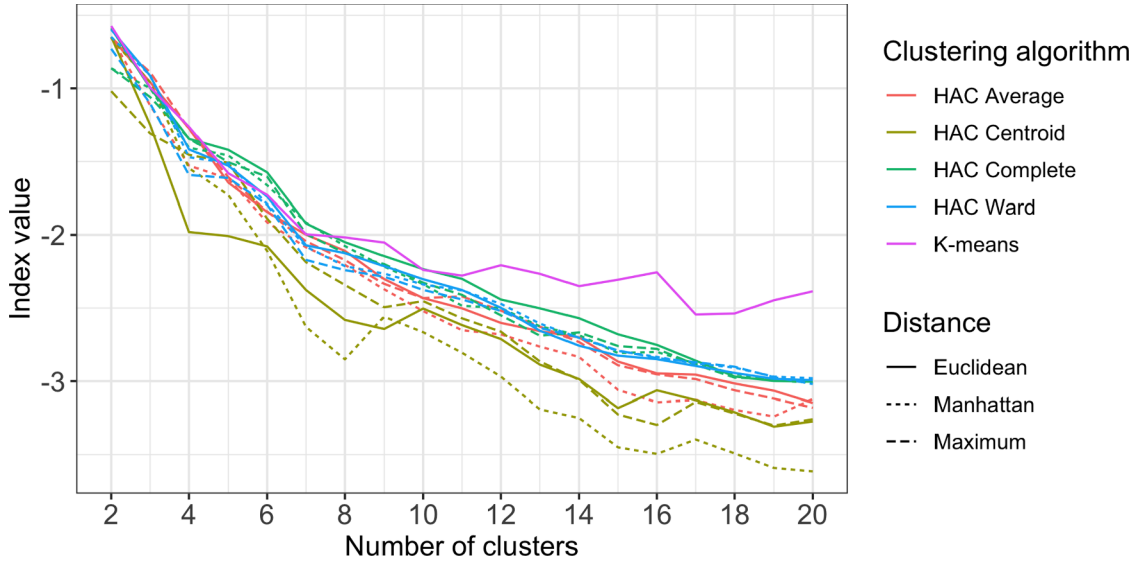


Fig. 6. Gap statistic values across 13 distance-algorithm combinations, evaluated for 2–20 clusters.

[84]. For this reason, the performance criteria sensitivity and specificity are commonly used for assessing classification performance. Sensitivity (SNS) is a measure of the classifier's ability to correctly identify positive labels. It is a ratio of the number of true positives (TP) to the sum of true positives and false negatives (FN), as in

$$\text{SNS} = \frac{TP}{TP + FN}. \tag{12}$$

Specificity (SPC) is a measure of the classifiers ability to correctly identify negative labels, or the ratio between the number of true negatives (TN) and the sum of the TN and false positives (FP). It is given by

$$\text{SPC} = \frac{TN}{TN + FP}. \tag{13}$$

The ROC curve is a convenient way to summarize the relationship between these two metrics to illustrate the trade-off between a classifier's true positive rate (TPR) and false positive rate (FPR), where the TPR is the same as the SNS, and the FPR can be written as $1 - \text{SPC}$. These metrics are computed given the classifier's predicted probabilities for each class across a number of different classification thresholds, and the results are plotted to show how the TPR and FPR are related. The AUC is a scale-invariant and classification-threshold-invariant metric that can be used to compare the general performance of a classifier by computing the total area under the plotted ROC curve. The AUC ranges in value from 0 to 1, with values closer to 1 indicating perfect classification and values equal to 0.5 or below indicating poor classifier performance.

For the logistic regression model with elastic net regularization, a sequence of 100 values for λ are evaluated to determine the overall complexity of the model [75]. A sequence of 10 α values are specified during the training procedure to find the best mix of ridge and lasso penalties. The hyperparameter to be tuned in the random forests model, the *mtry* parameter, is tuned across a sequence of values below and above $\sqrt{p}$ [72]. In both models, the hyperparameters are tuned such that the final selected model maximizes the AUC.

After selecting final models, we evaluate their performance on the test set and compute ROC curves and AUC values for each.

4.4. Software

We conduct all of our analyses in the *R Statistics* environment [85]. We use the following packages: for data processing and manipulation we use *dplyr* [86], and for plotting we use *ggplot2* [87]. Our comparative cluster analysis is carried out using the *NbClust* package [88]. For our classification models, we use the *caret* [89] and *caretEnsemble* [90] packages, which incorporate both the *glmnet* package [77] for regularization and the *ranger* package [91] for random forests. Finally, we use the *pROC* package [92] to compute and measure classifier performance.

5. Results

This section first presents results of the cluster analysis, including the optimal number of clusters selected and visualizations of these. Second, it presents results of the classification analysis, including the predictive power of activity data and key activities associated with varying peak-period electricity consumption patterns.

5.1. Cluster analysis results

Across all distance-algorithm combinations (including both k-means as well as hierarchical clustering with five different linkage methods), we evaluate results with several CVIs for 2 to 20 clusters. We find that of the CVIs with which we validate results, the gap statistic [93] shows the least variability of index scores across the different distance-algorithm methods tested, suggesting it is robust to numerous clustering approaches. Index scores across all six CVIs are shown in the appendix (Fig. A.2).

The gap statistic is a statistical testing method that compares the change in within-cluster dispersion given different numbers of clusters k with that expected under a reference null distribution of the data. The optimal number of clusters is that which maximizes the gap statistic (meaning the clustering structure is most different from a random uniform distribution).

We use the gap statistic to select the optimal number of clusters and assign final cluster membership. Fig. 6 presents values of the gap statistic for $k = 2, \ldots, 20$. It shows that across the 13 distance-algorithm combinations, 2 clusters is returned as optimal in all cases.

We assign load profiles to the cluster (1 or 2) each is most frequently clustered in across these 13 methods. This approach avoids the need to select a specific distance-algorithm combination to assign final cluster membership, especially since the index performance across most of the methods for $k = 2$ is similar (see Fig. 6).

Fig. 7 plots all load profiles along with the mean value at each hour for the two clusters. From left to right, we present visualizations of each

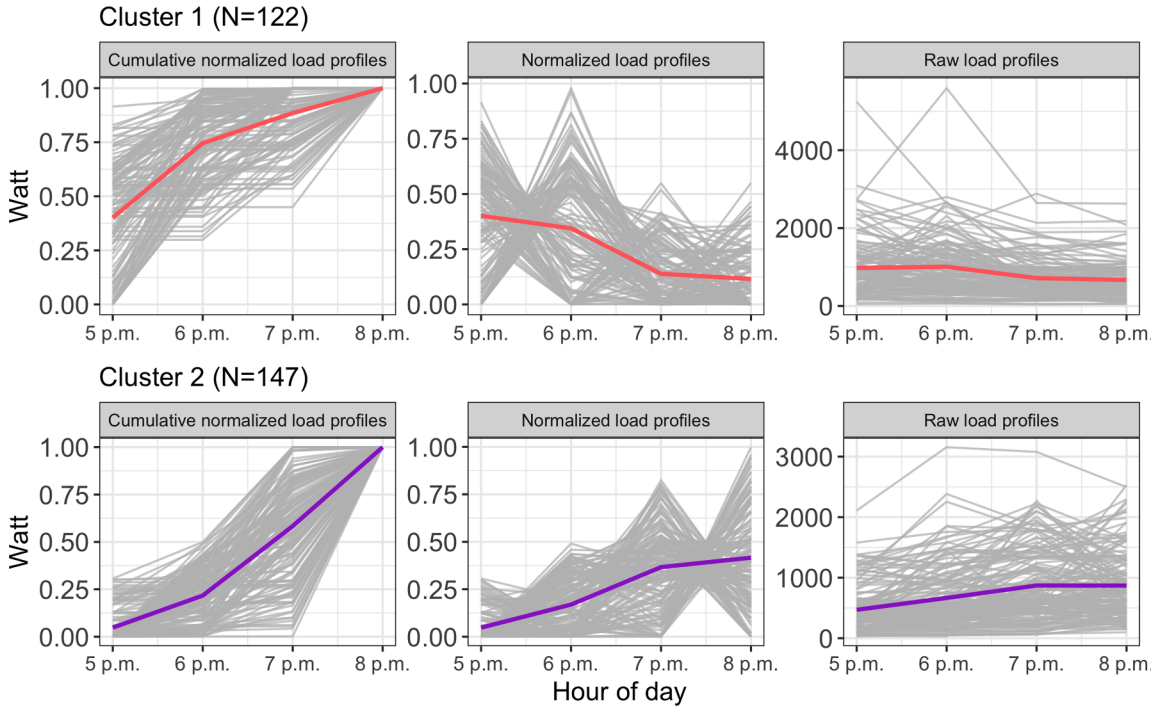


Fig. 7. Cluster analysis results for $K = 2$, where plots show member load profiles and cluster averages for cumulative, normalized, and raw data.

cluster and the cluster average for the normalized cumulative load profiles (the data on which the cluster analysis is performed), as well as the normalized clusters and the raw data. To highlight the methodological advantage of clustering cumulative rather than raw load profiles, we compute results for a counterfactual case, where the clustering is applied to de-minned profiles rather than cumulative de-minned profiles. This plot is included in the appendix (Fig. A.3) and shows how the cluster analysis is dominated by the magnitude of each profile's usage at 5 p.m. rather than its full shape throughout the evening period.

As shown in the top panel of Fig. 7, the first cluster consists of just under half the sample, and the average profile shape is characterized by higher discretionary usage in the earlier hours of the evening at 5 p.m. and then declining consumption through the peak period. The second cluster is slightly larger in sample size, and its average profile shape is characterized by lower usage at 5 p.m. but increased consumption through the peak period until 7 p.m., after which usage remains stable.

In Fig. 8, we show both cluster averages along with all load profiles on the same plot. Comparing the two, cluster 1's hourly usage of 978 W is nearly twice as high as cluster 2's (469 W) at 5 p.m. and around 1.5 times higher at 6 p.m. The cluster averages cross between 6 and 7 p.m., with Cluster 2's usage remaining around 150 W higher than Cluster 1's from 7 to 9 p.m.

The clustering approach demonstrated here fulfills the segmentation goal of identifying distinct clusters in terms of peak-period electric load profile shape. The results are especially useful for separating load profiles by the timing of peak demand, as cluster 1 should clearly be prioritized for interventions to reduce or shift demand during the early hours of the evening peak given its comparatively higher demand. Understanding the activity patterns driving these load profile groupings can help inform these efforts. For this reason, we next turn to the task of predicting cluster membership using occupant activity data.

5.2. Classification results

In both the logistic regression and the random forests models, we set up the prediction task as a binary classification problem, where the

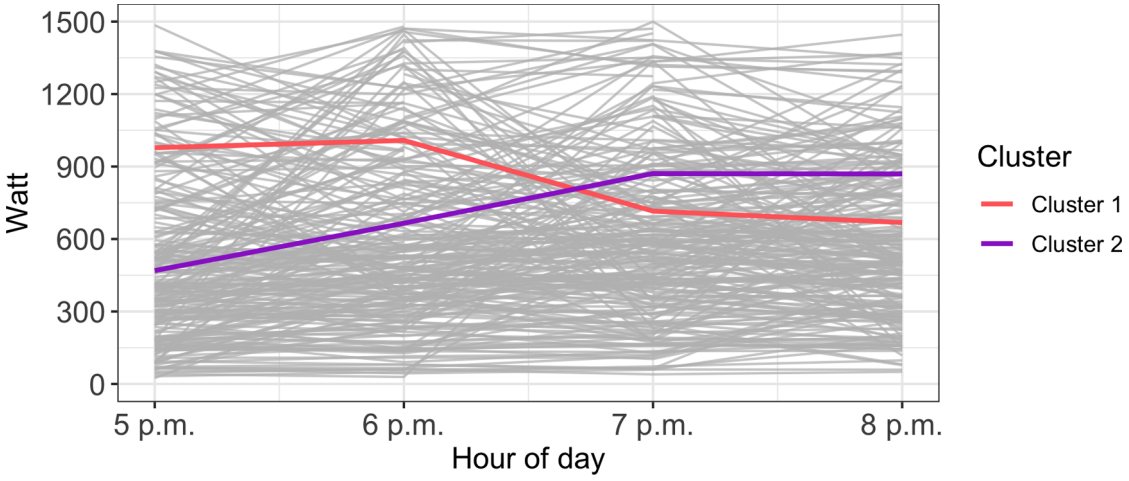


Fig. 8. Raw load profiles plotted with cluster averages for $K = 2$.

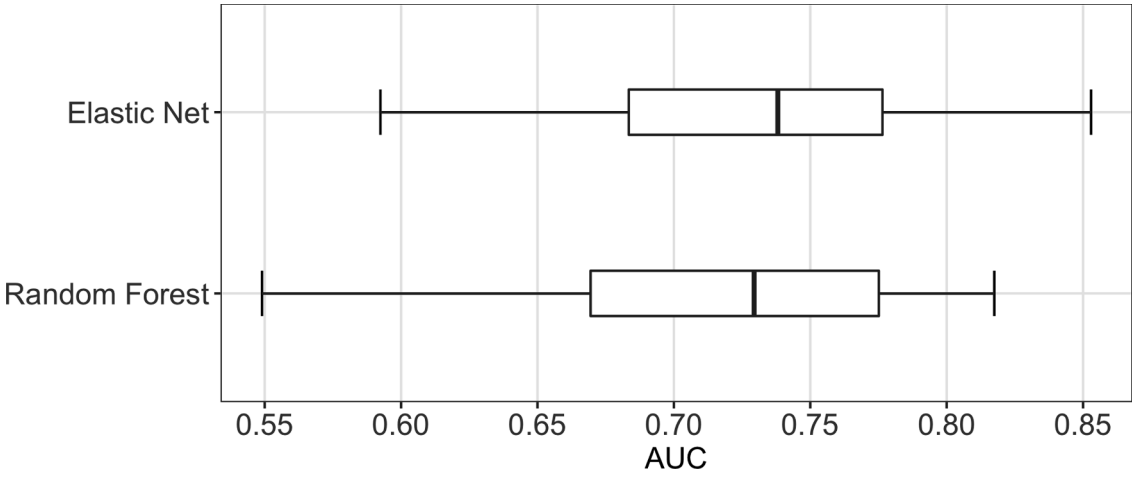


Fig. 9. Box-and-whisker plot of training AUC across 25 cross-validation resamples.

reference or base response category is membership in cluster 1, the early-peaking cluster. The models thus aim to predict the likelihood a load profile will be classified in the later-peaking cluster given the activities reported by occupants from 5 to 9 p.m.

In this section, we first present results related to model performance assessment and second on the predictive power of specific activities in the models.

5.2.1. Model performance assessment

We train our models on 70% of the data and use 5-fold repeated cross-validation to tune model hyperparameters and estimate test set error using AUC as the performance measure. The training results show that the best elastic net model is found with $\alpha = 0.22$ and $\lambda = 0.21$ and has an average cross-validation AUC of 0.73. The best random forests model is found with *mtry* = 5 and has an average cross-validation AUC of 0.72.

To assess the variability of these performance measures for the different samples taken during the cross-validation procedure, we use *Caret*'s built-in *resamples* function to compute the AUC metric across the 25 resamples (5 cross-validation folds with 5 repeats). Fig. 9 presents a box-and-whisker plot of the variability in model training AUC for these 25 cross-validation samples.

Fig. 9 shows that the elastic net model has a smaller interquartile range for training AUC than the random forests model, though there are not large performance differences between the models on the training data. Both do have relatively high overall ranges, however, suggesting the results are somewhat variable.

Next, we select the model with the tuning parameters that give the highest average cross-validated AUC in each case and use these to make predictions on the test set. We calculate ROC curves for each model's performance in predicting cluster membership for the test data. Fig. 10 shows these curves and AUC values for each model. In both cases, the model performance declines slightly on the test data, which is typical in most model prediction contexts [94]. The random forests model has a slightly higher test set AUC than the elastic net model, indicating that the elastic net model may have overfit the training data more than the random forests model.

5.2.2. Key activity variables

We inspect which activity predictors are influential in both classification models. For the elastic net model, we examine all non-zero coefficients that are selected, and for the random forests model, we assess variable importance using average impurity measured with the Gini index, as explained in Section 4.3.3. Table 3 presents predictors with non-zero coefficients in the elastic net model, including both their log(*odds*) (β) and exponentiated coefficients (Exp[β]). These are ordered

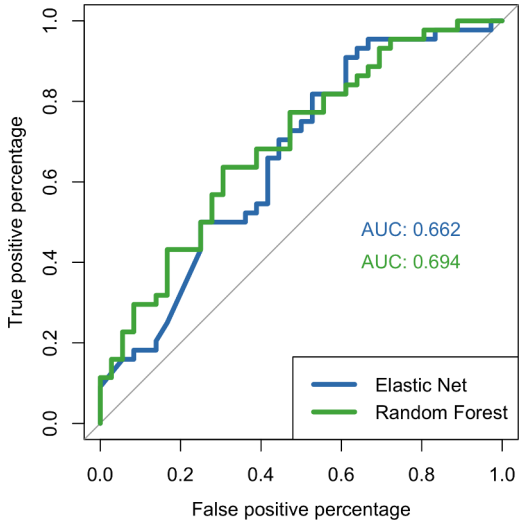


Fig. 10. ROC curves and AUC values for elastic net and random forests performance on the test set.

by hour and magnitude of β.

The elastic net regularization technique delivers substantial sparsity for the logistic regression model, selecting 22 variables from the set of 177 candidate predictors. This result highlights that many of the hour-activity variables, perhaps owing to their low frequency, could be excluded from the model without sacrificing predictive performance. Elastic net regularization is especially useful in this context, where $p \approx n$ after the testing data is held out. The technique helps balance the interests of including all hour-activity predictors as candidates for predicting peak-period electricity consumption patterns while also selecting a more interpretable model.

Given that the base category in the model is membership in the earlier-peaking cluster, negative β coefficients signal a decrease in the likelihood that a load profile will belong to the later-peaking cluster or, similarly, an increased likelihood it will belong to the earlier-peaking cluster. For instance, the predictor '5–6 p.m. Eating hot meal' has $\beta = -0.13$ and Exp[β] = 0.878. This means that for each additional time this activity is reported in a given household, its likelihood of being classified in the later-peaking cluster decreases by 12%. In the case of the '5–6 p.m. Laundry' activity, each additional time it is reported decreases the likelihood a household will be classified in the later-peaking cluster by 23.5%.

As Table 3 shows, nearly all of the earlier (5–6 or 6–7 p.m.) selected predictors have negative coefficients, whereas most of the activity

Table 3
Predictors with non-zero coefficients in final logistic model with elastic net regularization.

Hour (p.m.)	Activity group predictor	β	Exp(β)
5	Washing machine	− 0.399	0.671
5	Socializing	− 0.351	0.704
5	Laundry	− 0.268	0.765
5	Eat	− 0.151	0.860
5	Screen time	− 0.131	0.877
5	Eating hot meal	− 0.130	0.878
5	Tending domestic animals	0.107	1.113
5	Oven	− 0.105	0.900
5	Cook on hob	− 0.052	0.950
5	Other appliance	− 0.036	0.964
5	Dishwasher	0.024	1.024
6	Eating hot meal	− 0.337	0.714
6	Eat	− 0.307	0.735
6	Got home	0.286	1.331
6	Other appliance	0.178	1.194
6	Pet	− 0.162	0.850
6	Bath	− 0.149	0.861
7	Oven	0.152	1.164
7	Childcare	− 0.089	0.915
7	Screen time	− 0.009	0.991
8	Laundry	0.388	1.474
8	Oven	0.281	1.325
8	Eating hot meal	0.012	1.012
	Intercept	0.499	1.648

predictors reported later in the evening have positive coefficients. There are several exceptions, such as using the dishwasher from 5–6 p.m. or screen time from 7–8 p.m., but these generally have smaller coefficients.

In general, we find that meal-related activities are influential in the model, both in the case of earlier reports of meal activities showing strong associations with higher consumption patterns in the early evening and later reports of meal activities showing strong associations with patterns in the later evening. Other key activities include laundry and washing machine use, screen time, and socializing.

Turning to the random forests model, we plot normalized variable importance in Fig. 11, where the most important predictor is given a score of 1. We plot only the top 20 predictors for brevity.

As seen in the plot, many of the same variables that were influential in the elastic net model also have high variable importance scores in the random forests model. This is especially the case for the eating and meal-related predictors, screen time, and socializing.

An important difference between the linear elastic net model and the non-linear random forests model is that the predictors in the random forests model do not have coefficients and thus do not have a positive or negative sign and so cannot as easily be associated with increased probability of being in the early- or late-peaking cluster. Their importance is determined by how well they are able to split the predictor space. This lack of interpretability is one of the drawbacks of the random forests model, but considering many of the variables with high importance scores in this model are those selected in the elastic net model, it is reasonable to assume the direction of prediction is similar between the two. The random forests model, then, can provide further support to the findings on which activities are most influential for predicting peak-period electricity consumption patterns.

6. Discussion and conclusions

This paper presents an integrated analysis to cluster a sample of 269 peak-period electric load profiles from homes in the UK. We introduce a novel solution to capture temporal variations in consumption patterns by clustering cumulative load profiles. We show a new technique to appropriately use Euclidean distances with time-series data in order to cluster profiles based on their full shape during peak evening hours. This approach is beneficial for both its conceptual simplicity and its ease of implementation, and it does not appear to have been used previously for load profile segmentation.

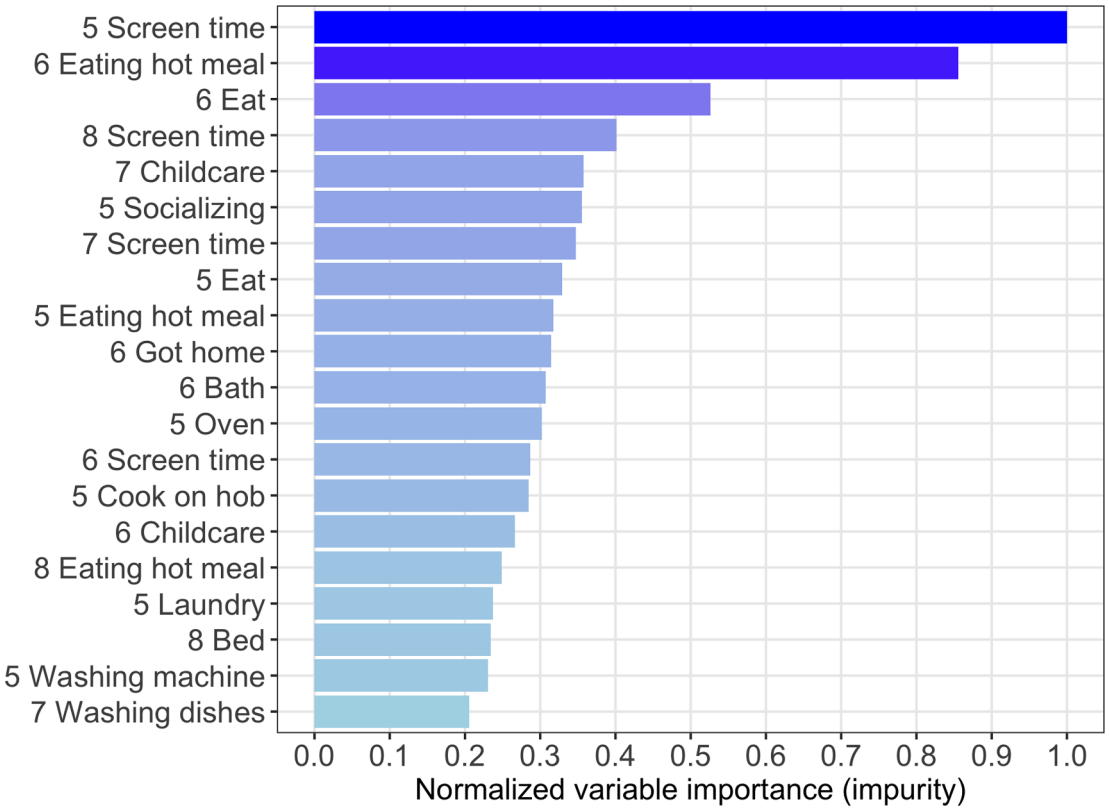


Fig. 11. Normalized variable importance as determined by the Gini impurity method for the random forests model.

We have shown how activity data can help to predict membership in early and late-peaking clusters. We introduce two classification methods that are appropriate for the predictive modeling task given analytical challenges inherent to this classification problem. The first is logistic regression with elastic net regularization and the second is random forests. Both are used to evaluate the predictive power of activities on held out test data, as well as to identify key activity variables for predicting cluster membership.

The two classification algorithms used to predict cluster membership from occupant-reported activities show test set AUC values of 0.66–0.69. These scores are not indicative of highly accurate predictive models, but they also clearly outperform a null classification model, which would have an AUC of 0.5. This finding suggests that activity data as it is incorporated in this analysis shows some promise for predicting patterns of electricity consumption during peak hours. To the authors' knowledge, this paper presents the first attempt to show how detailed time-use data can be used in a cluster-and-classification approach to better understand the activities driving electricity consumption patterns in the residential sector.

The identification of key activities driving evening demand is important given the complexity of the modeling task, for which simpler regression/classification techniques fail to give accurate and stable results. Both models confirm the importance of certain activities for predicting the shape of a household's evening weekday demand. The strongest link is found between food preparation and meal-time activities, especially cooking with energy-intensive appliances and eating hot meals. Other activities, including screen time, laundry, and socializing are shown to be important predictors, as well.

These empirical findings are useful for better understanding the potential for activity-led peak demand reduction and demand response in households [5]. 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 a greater response. For instance, our results suggest that the households that contribute the most to UK system-wide peak demand are those 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. Design of targeted interventions can benefit from taking this knowledge into account.

Similarly, deeper insight into electricity demand patterns in homes and their relationship to occupant activities can inform the design of more suitable electricity tariffs. Assessing household responses to time-varying electricity tariffs is an active area of research, but there is still little agreement on how adoption of such tariffs influences electricity use patterns [95]. Previous analyses of a smaller sample of households from the METER study provide some evidence that electricity tariff choice is associated with different patterns of daily electricity use. Satre-Meloy et al. [96] found that households enrolled in a renewable electricity tariff have lower average daily electricity consumption, as well as lower consumption during evening peak hours, than those on standard tariffs. The study also found that enrollment in an Economy 7 or 10 tariff, both of which charge lower prices for electricity use during nighttime hours, is linked to slightly higher nighttime electricity use. This evidence is based on a small number of households, however, as the share of those on either a renewable or Economy 7/10 tariff is less than 25%.

These initial findings are worth exploring in further detail with a larger sample of households enrolled in time-varying electricity tariffs. The implications of tariff selection on peak-period electricity usage patterns are especially important to understand given that time-varying tariffs are likely to become more common in the future with further deployment of smart metering infrastructure [97]. Furthermore, as most of the research on household responses to dynamic electricity prices has focused on how much of a peak reduction is achievable, the data collected in this study can provide insight into what activities are shifted, curtailed, or substituted to actually deliver those reductions. For instance, understanding whether or not households on electricity tariffs that charge higher prices during peak hours are more likely to shift high-consumption activities, such as laundry, cooking, or screen time activities—all of which were shown in this study to be predictive of evening peak-period consumption patterns—can provide unique understanding on which types of activities may be more or less price-responsive. This knowledge could further be used to inform tariff design for different classes of customers, identify constraints to demand response in households, or enable better integration of demand-side resources in the electricity system [18].

Some limitations are present in the data and analysis. The sample size for this study is small, non-representative, and it consists of households that participate across different seasons and days of the week, excluding weekends (though most of our sample participated during the winter season). Day of week and seasonal differences are no doubt important for understanding the temporal aspects of activities and electricity consumption, and future larger samples could be segmented to explore such variations. We also expect biases are present in our sample's activity reporting and in its socio-demographic makeup. Participant self-selection is a further source of bias.

Regarding the statistical methods we use, the predictive power demonstrated in the classification models is somewhat limited, though we expect this might improve with access to larger training data sets. Training AUC across cross-validation resamples also has high variability. Cluster analysis is typically applied to larger data sets, and for many of the CVIs used to evaluate our clustering results, the index scores are highly variable for different numbers of clusters. We use the Gap index for final cluster assignment because it shows less variability than the other CVIs, but our exploratory analysis of cluster results across different methods and CVIs shows that final segmentation of the load profiles is dependent on the clustering algorithm and distance metric chosen.

We expect the approach demonstrated in this paper to yield clearer insights when applied to a larger, more representative sample, and for this reason, data collection is ongoing. Future work should further assess the performance of both cluster and classification analyses presented in this paper. For the cluster analysis approach, we advocate for further comparisons of cumulative load profile clustering with other time-series-specific clustering approaches. For the classification analysis, we encourage similar analyses of the predictive power of time-use activity data. While we restrict our focus to predicting evening peak-period consumption patterns, a similar approach could be applied to different times of day or to longer periods. The approach to aggregate and represent activities in the models should be tested in future analyses, especially the use of different time windows for aggregation (for instance, counting all activities reported every half hour rather than hour). Further work should also explore how results change if the modeling approach is set up as a regression rather than classification problem, especially if the key concern is predicting drivers of peak demand.

As one key objective of this analysis is to identify activities that could be specifically targeted in demand response programs, future work should also aim to test whether and how much these kinds of targeted interventions can increase reductions in usage during peak hours. Collecting data on the kinds of activities that are shifted during demand response interventions provides numerous avenues for further research, such as exploring whether occupant responses to interventions are sustainable over longer time periods or understanding the unintended consequences that might result from these interventions. At

present, utilities receive little feedback of this nature, and there thus remain substantial opportunities to better inform demand response interventions. 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.

Data availability

The data used in this study are publicly available from the UK Data Service [98].

Appendix A

Table A.1.
See Figs. A.1–A.3.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge funding for this research from the Engineering and Physical Sciences Research Council (EPSRC) [EP/M024652/1]. The authors declare no competing interests.

Table A.1
Cluster validity indices (CVIs) for determining optimal number of clusters.

Index	Equation	Optimal number of clusters
Caliński and Harabasz [99]	$\mathrm{CH}(q) = \frac{\mathrm{trace}(B_q)/(q-1)}{\mathrm{trace}(W_q)/(n-q)}$ where: • W_q is the within-group dispersion matrix for data clustered into q clusters, • B_q is the between-group dispersion matrix for data clustered into q clusters, • q is the number of clusters, • n is the number of observations	Maximum of index
C-index [100]	$\mathrm{Cindex} = \frac{S_w - S_{\min}}{S_{\max} - S_{\min}}$, $S_{\min} \neq S_{\max}$, $\mathrm{Cindex} \in (0, 1)$ where: • S_w is the sum of within-cluster distances $S_w = \sum_{k=1}^{q} \sum_{i<j}^{i,j \in C_k} d(x_i, x_j)$ • S_{min} is the sum of the N_w smallest distances between all the pairs of points in the entire data set (there are N_t such pairs), • S_{max} is the sum of the N_w largest distances between all the pairs of points in the entire data set (there are N_t such pairs)	Minimum of index
Davies-Bouldin [101]	$\mathrm{DB}(q) = \frac{1}{q} \sum_{k=1}^{q} \max_{k \neq l} \left(\frac{\delta_k + \delta_l}{d_{kl}}\right)$ where: • $k, l = 1, \ldots, q$ is the cluster number, • d_{kl} is the distance between centroids of clusters C_k and C_l, • δ_k is the dispersion measure of a cluster C_k	Minimum of index
Duda [102]	$\mathrm{Duda} = \frac{Je(2)}{Je(1)} = \frac{W_k + W_l}{W_m}$ where: • $J_e(2)$ is the sum of squared errors within clusters when the data are partitioned into two clusters, • $J_e(1)$ gives the squared errors when only one cluster is present. • It is assumed clusters C_k and C_l are merged to form C_m	Smallest number of clusters such that index > critical value [see 103]
Gap statistic [93]	$\mathrm{Gap}(q) = \frac{1}{B} \sum_{b=1}^{B} \log W_{qb} - \log W_q$ where: • B is the number of reference data sets generated using uniform prescription, • W_{qb} is the within-dispersion matrix defined as in Hartigan [104]	Smallest number of clusters such that critical value $\geqslant 0$
Silhouette [105]	$\mathrm{Silhouette} = \frac{\sum_{i=1}^{n} S(i)}{n}$, $\mathrm{Silhouette} \in [-1, 1]$ where: • $S(i) = \frac{b(i) - a(i)}{\max\{a(i); b(i)\}}$ • $a(i)$ is average dissimilarity of the ith object to all other objects of cluster C_r • $b(i) = \min d_{iC_s}$ • d_{iC_s} is the average dissimilarity of the ith object to all objects of cluster C_s	Maximum of index

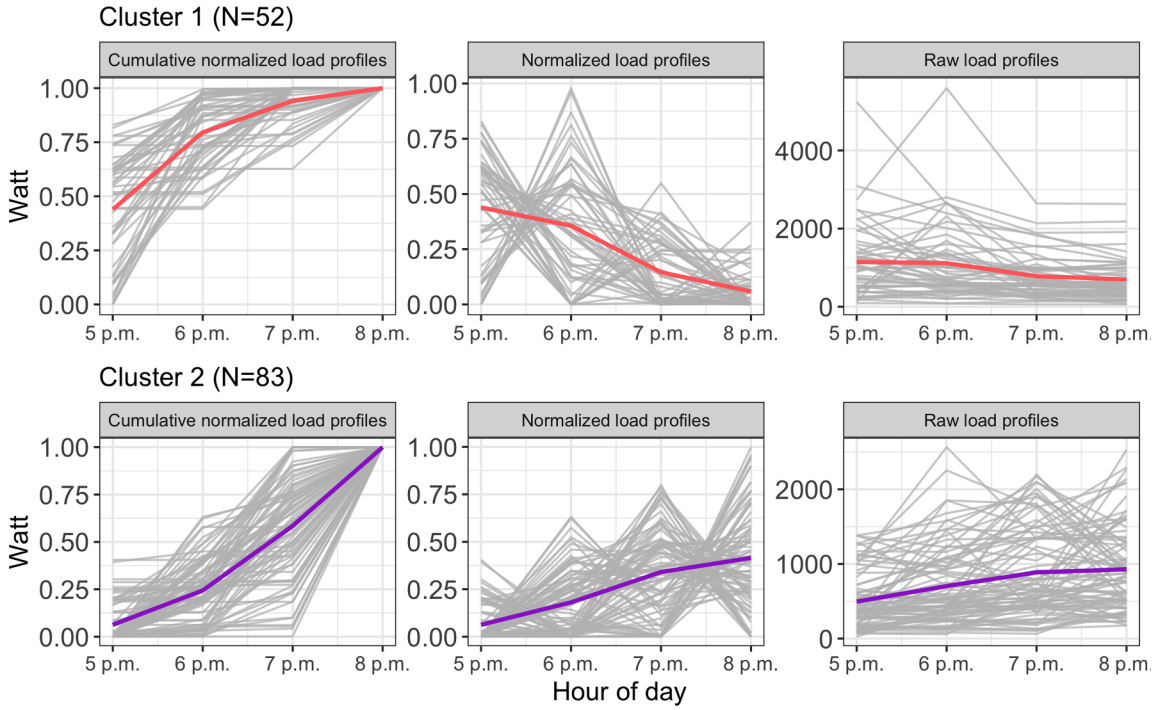


Fig. A.1. Clustering results using only load profiles from each household's first day of the study.

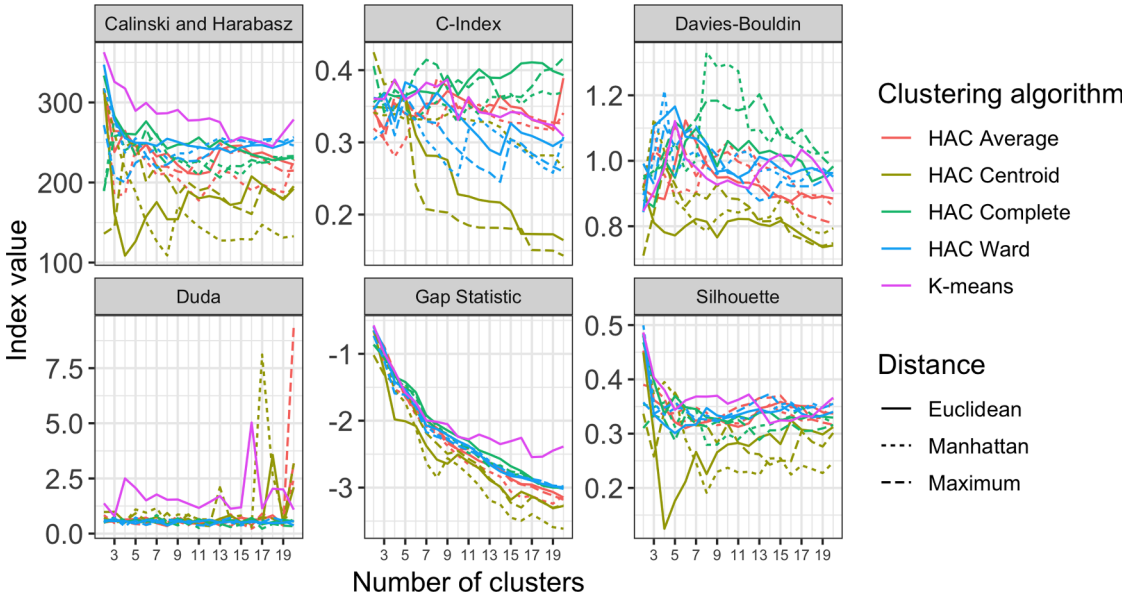


Fig. A.2. Results for cluster analysis evaluated with six CVIs for 2–20 clusters to show index score variability across different method combinations.

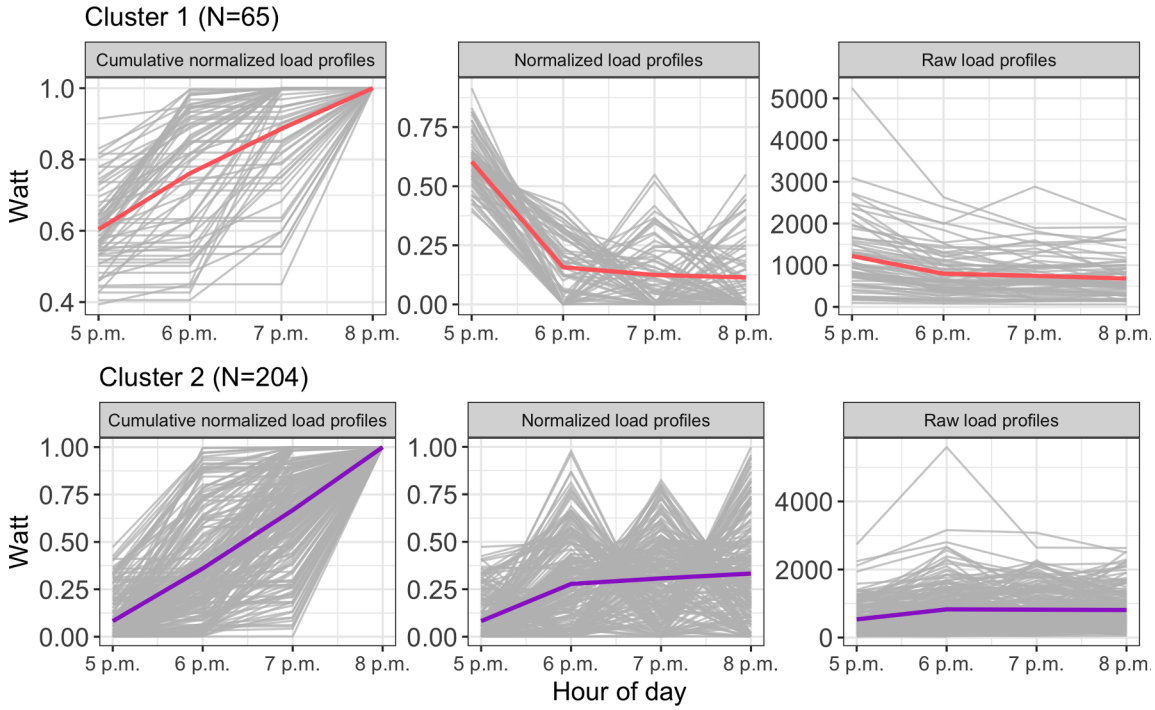


Fig. A.3. Results for cluster analysis applied to normalized load profiles (middle column plots) rather than cumulative normalized load profiles (left column plots).

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.apenergy.2019.114246.

References

[1] BEIS, 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; 2018. p. 111–53.

[2] Ofgem, Demand Side Reponse. A Discussion Paper, Tech. Rep., Office of Gas and Electricity Markets, London, United Kingdom; 2010.

[3] McLoughlin F, Duffy A, Conlon M. Evaluation of time series techniques to characterise domestic electricity demand. Energy 2013;50:120–30. https://doi.org/10.1016/j.energy.2012.11.048. ISSN 0360-5442.

[4] National Infrastructure Commission, Smart Power, Tech. Rep. London, United Kingdom; 2016.

[5] Grunewald P, Diakonova M. Flexibility, dynamism and diversity in energy supply and demand: a critical review. Energy Res Soc Sci 2018;38:58–66. https://doi.org/10.1016/j.erss.2018.01.014. ISSN 22146296.

[6] Satre-Meloy A, Langevin J. Assessing the time-sensitive impacts of energy efficiency and flexibility in the US building sector. Environ Res Lett 2019;14(12). https://doi.org/10.1088/1748-9326/ab512e.

[7] Satre-Meloy A. Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models. Energy 2019;174:148–68. https://doi.org/10.1016/j.energy.2019.01.157. ISSN 0360-5442.

[8] Albert A, Rajagopal R. Smart meter driven segmentation: what your consumption says about you. IEEE Trans Power Syst 2013;28(4):4019–30. https://doi.org/10.1109/TPWRS.2013.2266122. ISSN 0885-8950.

[9] Granell R, Axon CJ, Wallom DCH. Clustering disaggregated load profiles using a Dirichlet process mixture model. Energy Convers Manage 2015;92:507–16. https://doi.org/10.1016/j.enconman.2014.12.080. ISSN 0196-8904.

[10] Kwac J, Flora J, Rajagopal R. Lifestyle segmentation based on energy consumption data. IEEE Trans Smart Grid 2018;9(4):2409–18. https://doi.org/10.1109/TSG.2016.2611600. ISSN 1949-3053.

[11] McLoughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl Energy 2015;141:190–9. https://doi.org/10.1016/j.apenergy.2014.12.039. ISSN 03062619.

[12] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. Appl Energy 2014;135:461–71. https://doi.org/10.1016/j.apenergy.2014.08.111. ISSN 03062619.

[13] Viegas JL, Vieira SM, Melício R, Mendes V, Sousa JM. Classification of new electricity customers based on surveys and smart metering data. Energy 2016;107:804–17. https://doi.org/10.1016/j.energy.2016.04.065. ISSN 03605442.

[14] Liao TW. Clustering of time series data—a survey. Pattern Recogn 2005;38(11):1857–74. https://doi.org/10.1016/j.patcog.2005.01.025. ISSN 00313203.

[15] Beckel C, Sadamori L, Staake T, Santini S. Revealing household characteristics from smart meter data. Energy 2014;78:397–410. https://doi.org/10.1016/j.energy.2014.10.025. ISSN 0360-5442.

[16] Räsänen T, Voukantsis D, Niska H, Karatzas K, Kolehmainen M. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Appl Energy 2010;87(11):3538–45. https://doi.org/10.1016/j.apenergy.2010.05.015. ISSN 0306-2619.

[17] Stephen B, Mutanen AJ, Galloway S, Burt G, Jarventausta P. Enhanced load profiling for residential network customers. IEEE Trans Power Delivery 2014;29(1):88–96. https://doi.org/10.1109/TPWRD.2013.2287032. ISSN 0885-8977, 1937-4208.

[18] Flath C, Nicolay D, Conte T, van Dinther C, Filipova-Neumann L. Cluster analysis of smart metering data. Bus Inform Syst Eng 2012;4(1):31–9. https://doi.org/10.1007/s12599-011-0201-5. ISSN 1867-0202.

[19] Cao H-A, Beckel C, Staake T. Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns. In: IECON 2013–39th annual conference of the IEEE industrial electronics society. Vienna, Austria: Institute of Electrical and Electronics Engineers; 2013. p. 4733–38. doi: https://doi.org/10.1109/IECON.2013.6699900. ISBN 978-1-4799-0224-8.

[20] Kwac J, Flora J, Rajagopal R. Household energy consumption segmentation using hourly data. IEEE Trans Smart Grid 2014;5(1):420–30. https://doi.org/10.1109/TSG.2013.2278477. ISSN 1949-3053.

[21] Hsu D. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. Appl Energy 2015;160:153–63. https://doi.org/10.1016/j.apenergy.2015.08.126. ISSN 03062619.

[22] Iglesias F, Kastner W. Analysis of similarity measures in times series clustering for the discovery of building energy patterns. Energies 2013;6(2):579–97. https://doi.org/10.3390/en6020579.

[23] Piao M, Shon HS, Lee JY, Ryu KH. Subspace projection method based clustering analysis in load profiling. IEEE Trans Power Syst 2014;29(6):2628–35. https://doi.org/10.1109/TPWRS.2014.2309697. ISSN 0885-8950.

[24] du Toit J, Davimes R, Mohamed A, Patel K, Nye JM. Customer segmentation using unsupervised learning on daily energy load profiles. J Adv Inform Technol 2016;7(2):69–75. https://doi.org/10.12720/jait.7.2.69-75. ISSN 17982340.

[25] Smith BA, Wong J, Rajagopal R. A simple way to use interval data to segment residential customers for energy efficiency and demand response program targeting. In: Proceedings of the 2012 ACEEE summer study on energy efficiency in buildings. Pacific Grove, CA: American Council for an Energy Efficient Economy, vol. 13; 2012.

[26] Haben S, Singleton C, Grindrod P. Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Trans Smart Grid 2016;7(1):136–44. https://doi.org/10.1109/TSG.2015.2409786. ISSN 1949-3053, 1949-3061.

[27] Teeraratkul T, O'Neill D, Lall S. Shape-based approach to household electric load curve clustering and prediction. IEEE Trans Smart Grid 2018;9(5):5196–206. https://doi.org/10.1109/TSG.2017.2683461. ISSN 1949-3053.

[28] Wang Y, Chen Q, Kang C, Zhang M, Wang K, Zhao Y. Load profiling and its application to demand response: a review. Tsinghua Sci Technol 2015;20(2):117–29. https://doi.org/10.1109/TST.2015.7085625. ISSN 1007-0214.

[29] Zhou K-L, Yang S-L, Shen C. A review of electric load classification in smart grid environment. Renew Sustain Energy Rev 2013;24:103–10. https://doi.org/10.1016/j.rser.2013.03.023. ISSN 1364-0321.

[30] Figueiredo V, Rodrigues F, Vale Z, Gouveia JB. An electric energy consumer characterization framework based on data mining techniques. IEEE Trans Power Syst 2005;20(2):596–602. https://doi.org/10.1109/TPWRS.2005.846234. ISSN 0885-8950.

[31] Azaza M, Wallin F. Smart meter data clustering using consumption indicators: responsibility factor and consumption variability. In: Proceedings of the 9th international conference on applied energy, vol. 142. Energy Procedia, Cardiff, United Kingdom; 2017. p. 2236–42. doi: https://doi.org/10.1016/j.egypro.2017.12.624.

[32] Chicco G, Napoli R, Piglione FP. Comparisons among clustering techniques for electricity customer classification. IEEE Trans Power Syst 2006;21(2):933–40. https://doi.org/10.1109/TPWRS.2006.873122. ISSN 0885-8950.

[33] Dent I, Aickelin U, Rodden T. Application of a Clustering Framework to U.K Domestic Electricity Data. In: Proceedings of the 11th annual workshop on computational intelligence. Manchester, United Kingdom: UK Computational Intelligence; 2011. p. 161–6.

[34] Rodrigues F, Duarte J, Figueiredo V, Vale Z, Cordeiro M. A comparative analysis of clustering algorithms applied to load profiling. In: Perner P, Rosenfeld A, editors. Machine learning and data mining in pattern recognition. Springer-Verlag; 2003. p. 73–85. ISBN 978-3-540-45065-8.

[35] Gouveia JP, Seixas J. Unraveling electricity consumption profiles in households through clusters: combining smart meters and door-to-door surveys. Energy Build 2016;116:666–76. https://doi.org/10.1016/j.enbuild.2016.01.043. ISSN 0378-7788.

[36] Granell R, Axon CJ, Wallom DCH. Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles. IEEE Trans Power Syst 2015;30(6):3217–24. https://doi.org/10.1109/TPWRS.2014.2377213. ISSN 0885-8950.

[37] Jin L, Lee D, Sim A, Borgeson S, Wu K, Spurlock CA, et al. Comparison of clustering techniques for residential energy behavior using smart meter data. In: Proceedings of the thirty-first AAAI conference on artifical intelligence, association for the advancement of artificial intelligence, San Francisco, CA; 2017.

[38] Kwac J, Tan C, Sintov N, Flora J, Rajagopal R. Utility customer segmentation based on smart meter data: empirical study. In: Proceedings of the 2013 IEEE international conference on smart grid communications (SmartGridComm). Vancouver, Canada: Institute of Electrical and Electronics Engineers; 2013. p. 720–5. doi: https://doi.org/10.1109/SmartGridComm.2013.6688044.

[39] Liu H, Yao Z, Eklund T, Back B. Electricity consumption time series profiling: a data mining application in energy industry. In: Perner P, editor. Advances in data mining. applications and theoretical aspects. Heidelberg: Springer, Berlin; 2012. p. 52–66. doi: https://doi.org/10.1007/978-3-642-31488-9_5. ISBN 978-3-642-31488-9.

[40] Panapakidis IP, Papadopoulos TA, Christoforidis GC, Papagiannis GK. Pattern recognition algorithms for electricity load curve analysis of buildings. Energy Build 2014;73:137–45. https://doi.org/10.1016/j.enbuild.2014.01.002. ISSN 0378-7788.

[41] Tsekouras GJ, Kotoulas PB, Tsirekis CD, Dialynas EN, Hatziargyriou ND. A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. Electr Power Syst Res 2008;78(9):1494–510. https://doi.org/10.1016/j.epsr.2008.01.010. ISSN 0378-7796.

[42] Xu S, Barbour E, González MC. Household segmentation by load shape and daily consumption. In: Proceedings of ACM SigKDD 2017 conference, Halifax, Nova Scotia; 2017. doi: 10.475/123_4.

[43] Zhang T, Zhang G, Lu J, Feng X, Yang W. A new index and classification approach for load pattern analysis of large electricity customers. IEEE Trans Power Syst 2012;27(1):153–60. https://doi.org/10.1109/TPWRS.2011.2167524. ISSN 0885-8950.

[44] Wei Y, Zhang X, Shi Y, Xia L, Pan S, Wu J, et al. A review of data-driven approaches for prediction and classification of building energy consumption. Renew Sustain Energy Rev 2018;82:1027–47. https://doi.org/10.1016/j.rser.2017.09.108. ISSN 1364-0321.

[45] Swan LG, Ugursal VI. Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. Renew Sustain Energy Rev 2009;13(8):1819–35. https://doi.org/10.1016/j.rser.2008.09.033. ISSN 13640321.

[46] Parti M, Parti C. The total and appliance-specific conditional demand for electricity in the household sector. Bell J Econ 1980;11(1):309–21. https://doi.org/10.2307/3003415. ISSN 0361-915X.

[47] Zhao H-X, Magoulès F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012;16(6):3586–92. https://doi.org/10.1016/j.rser.2012.02.049. ISSN 1364-0321.

[48] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. Renew Sustain Energy Rev 2018;81:1192–205. https://doi.org/10.1016/j.rser.2017.04.095. ISSN 1364-0321.

[49] Edwards RE, New J, Parker LE. Predicting future hourly residential electrical consumption: a machine learning case study. Energy Build 2012;49:591–603. https://doi.org/10.1016/j.enbuild.2012.03.010. ISSN 03787788.

[50] Tso GK, Yau KK. Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks. Energy 2007;32(9):1761–8. https://doi.org/10.1016/j.energy.2006.11.010. ISSN 03605442.

[51] Yu Z, Haghighat F, Fung BCM, Yoshino H. A decision tree method for building energy demand modeling. Energy Build 2010;42(10):1637–46. https://doi.org/10.1016/j.enbuild.2010.04.006. ISSN 0378-7788.

[52] Torriti J. A review of time use models of residential electricity demand. Renew Sustain Energy Rev 2014;37:265–72. https://doi.org/10.1016/j.rser.2014.05.034. ISSN 13640321.

[53] Gershuny J. Time-use studies: daily life and social change: full research report. ESRC end of award report, RES-000-23-0704-A. Swindon, United Kingdom: ESRC; 2008.

[54] U.S. Bureau of Labor Statistics, American Time Use Survey: Handbook of Methods, Tech. Rep., U.S. Department of Labor, Washington, D.C.; 2018.

[55] McKenna E, Thomson M. High-resolution stochastic integrated thermal-electrical domestic demand model. Appl Energy 2016;165:445–61. https://doi.org/10.1016/j.apenergy.2015.12.089. ISSN 03062619.

[56] Richardson I, Thomson M, Infield D, Clifford C. Domestic electricity use: a high-resolution energy demand model. Energy Build 2010;42(10):1878–87. https://doi.org/10.1016/j.enbuild.2010.05.023. ISSN 03787788.

[57] Torriti J, Hanna R, Anderson B, Yeboah G, Druckman A. Peak residential electricity demand and social practices: deriving flexibility and greenhouse gas intensities from time use and locational data. Indoor Built Environ 2015;24(7):891–912. https://doi.org/10.1177/1420326X15600776. ISSN 1420-326X.

[58] Widén J, Lundh M, Vassileva I, Dahlquist E, Ellegård K, Wäckelgård E. Constructing load profiles for household electricity and hot water from time-use data—modelling approach and validation. Energy Build 2009;41(7):753–68. https://doi.org/10.1016/j.enbuild.2009.02.013. ISSN 0378-7788.

[59] Widén J, Wäckelgård E. A high-resolution stochastic model of domestic activity patterns and electricity demand. Appl Energy 2010;87(6):1880–92. https://doi.org/10.1016/j.apenergy.2009.11.006. ISSN 0306-2619.

[60] De Lauretis S, Ghersi F, Cayla J-M. Energy consumption and activity patterns: an analysis extended to total time and energy use for French households. Appl Energy 2017;206:634–48. https://doi.org/10.1016/j.apenergy.2017.08.180. ISSN 03062619.

[61] Druckman A, Buck I, Hayward B, Jackson T. Time, gender and carbon: a study of the carbon implications of British adults' use of time. Ecol. Econ. 2012;84:153–63. https://doi.org/10.1016/j.ecolecon.2012.09.008. ISSN 0921-8009.

[62] Jalas M, Juntunen JK. Energy intensive lifestyles: time use, the activity patterns of consumers, and related energy demands in Finland. Ecol Econ 2015;113:51–9. https://doi.org/10.1016/j.ecolecon.2015.02.016. ISSN 09218009.

[63] Anderson B, Torriti J. Explaining shifts in U.K. electricity demand using time use data from 1974 to 2014. Energy Policy 1974;123(2018):544–57. https://doi.org/10.1016/j.enpol.2018.09.025. ISSN 0301-4215.

[64] Palmer J, Terry N, Kane T, Firth S, Hughes M, Pope P, et al. Further analysis of the household electricity use survey - electrical appliances at home: tuning in to energy saving. Tech. Rep. 475/09/2012. London, United Kingdom: Cambridge Architectural Research Limited, Elementenergy, and Loughborough University, 2013.

[65] McKenna E, Higginson S, Grunewald P, Darby SJ. Simulating residential demand response: improving socio-technical assumptions in activity-based models of energy demand. Energy Efficiency doi: https://doi.org/10.1007/s12053-017-9525-4. ISSN 1570-646X, 1570-6478.

[66] Grunewald P, Diakonova M. The electricity footprint of household activities-implications for demand models. Energy Build 2018;174:635–41. https://doi.org/10.1016/j.enbuild.2018.06.034. ISSN 03787788.

[67] Grünewald P, Layberry R. Measuring the relationship between time-use and electricity consumption. In: Proceedings of the 2015 ECEEE summer study on energy efficiency in buildings. European Council for an Energy Efficient Economy, Presqu'ı̈le de Giens, France; 2015. p. 2087–96.

[68] eurostat, Harmonised European Time Use Surveys: 2008 Guidelines, Office for Official Publications of the European Communities, Luxembourg; 2009. ISBN 978-92-79-07853-8, oCLC: 698893637.

[69] Grünewald P, Diakonova M. METER Data; 2019a. http://www.energy-use.org/docs/data/.

[70] Grünewald P, Diakonova M, Zilli D, Bernard J, Matousek A. 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. European Council for an Energy Efficient Economy, Presqu'ı̈le de Giens, France; 2017. p. 2085–93.

[71] Jin L, Spurlock A, Borgeson S, Fredman D, Hans L, Patel S, et al. Load shape clustering using residential smart meter data: a technical memorandum. Tech. Rep. Berkeley, CA: Lawrence Berkeley National Laboratory; 2016.

[72] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Springer series in statistics vol. 1. Berlin, Germany: Springer; 2001.

[73] Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 1984;71(1):1–10. https://doi.org/10.2307/2336390. ISSN 0006-3444.

[74] Farrar DE, Glauber RR. Multicollinearity in regression analysis: the problem revisited. Rev Econ Stat 1967;49(1):92. https://doi.org/10.2307/1937887. ISSN 00346535.

[75] Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso

and generalizations. Chapman & Hall/CRC; 2015. ISBN 978-1-4987-1216-3.
[76] Zou H, Hastie T. Regularization and variable selection via the elastic net. J Roy Stat Soc Ser B (Stat Methodol) 2005;67(2):301–20.
[77] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1.
[78] Taylor J, Tibshirani RJ. Statistical learning and selective inference. Proc Nat Acad Sci 2015;112(25):7629–34. https://doi.org/10.1073/pnas.1507583112. 1091–6490, ISSN 0027-8424.
[79] Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. Ann Stat 2014;42(2):413.
[80] Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the Lasso. Ann Stat 2016;44(3):907–27. https://doi.org/10.1214/15-AOS1371. ISSN 0090-5364, 2168-8966.
[81] Ma J, Cheng JC. Identifying the influential features on the regional energy use intensity of residential buildings based on random forests. Appl Energy 2016;183:193–201. https://doi.org/10.1016/j.apenergy.2016.08.096. ISSN 03062619.
[82] Wang Z, Wang Y, Zeng R, Srinivasan RS, Ahrentzen S. Random forest based hourly building energy prediction. Energy Build 2018;171:11–25. https://doi.org/10.1016/j.enbuild.2018.04.008. ISSN 0378-7788.
[83] Raileanu LE, Stoffel K. Theoretical comparison between the Gini index and information gain criteria. Ann Math Artif Intell 2004;41(1):77–93. https://doi.org/10.1023/B:AMAI.0000018580.96245.c6. ISSN 1573-7470.
[84] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inform Process Manage 2009;45(4):427–37. https://doi.org/10.1016/j.ipm.2009.03.002. ISSN 0306-4573.
[85] R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2017.
[86] Wickham H, François R, Henry L, Müller K. Dplyr: a grammar of data manipulation; 2019.
[87] Wickham H. Ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016.
[88] Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. J Stat Softw 61(6). doi: https://doi.org/10.18637/jss.v061.i06. ISSN 1548-7660.
[89] Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28(1):1–26. https://doi.org/10.18637/jss.v028.i05. ISSN 1548-7660.
[90] Deane-Mayer ZA, Knowles JE. caretEnsemble: Ensembles of Caret Models; 2016.
[91] Wright MN, Ziegler A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 2017;77(1):1–17. https://doi.org/10.18637/jss.v077.i01.
[92] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: Display and Analyze ROC Curves; 2019.
[93] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J Roy Stat Soc Ser B (Stat Methodol) 2001;63(2):411–23. https://doi.org/10.1111/1467-9868.00293.
[94] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer texts in statistics, vol. 103. New York, NY: Springer New York; 2013. doi: https://doi.org/10.1007/978-1-4614-7138-7. ISBN 978-1-4614-7137-0 978-1-4614-7138-7.
[95] Harding M, Sexton S. Household response to time-varying electricity prices. Annu Rev Resour Econ 2017;9(1):337–59. https://doi.org/10.1146/annurev-resource-100516-053437. ISSN 1941-1340, 1941-1359.
[96] Satre-Meloy A, Diakonova M, Grünewald P. Daily life and demand: an analysis of intra-day variations in residential electricity consumption with time-use data. Energ Effi 2019:1–26. https://doi.org/10.1007/s12053-019-09791-1. ISSN 1570-646X, 1570-6478.
[97] Nicolson ML, Fell MJ, Huebner GM. Consumer demand for time of use electricity tariffs: a systematized review of the empirical evidence. Renew Sustain Energy Rev 2018;97:276–89. https://doi.org/10.1016/j.rser.2018.08.040. ISSN 1364-0321.
[98] Grünewald P, Diakonova M. METER: U.K. Household Electricity and Activity Survey, 2016–2019: Secure Access. [Data Collection], UK Data Service SN: 8475. itemType: dataset. doi: https://doi.org/10.5255/UKDA-SN-8475-1.
[99] Caliński T, Harabasz J. A Dendrite method for cluster analysis. Commun Stat 1974;3(1):1–27. https://doi.org/10.1080/03610927408827101. ISSN 0090-3272.
[100] Hubert LJ, Levin JR. A general statistical framework for assessing categorical clustering in free recall, Psychol Bull 1976; 83(6):1072–80. ISSN 0033–2909.
[101] Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell PAMI-1 1979;2:224–7. doi: https://doi.org/10.1109/TPAMI.1979.4766909. ISSN 0162-8828.
[102] Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley-Blackwell; 1973. ISBN 978-0-471-22361-0.
[103] Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika 1985;50(2):159–79. https://doi.org/10.1007/BF02294245. ISSN 1860-0980.
[104] Hartigan JA. Clustering algorithms. Wiley series in probability and mathematical statistics. New York, NY: John Wiley & Sons Inc; 1975. ISBN 978-0-471-35645-5.
[105] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7. ISSN 0377-0427.