Electricity consumption pattern analysis beyond traditional clustering methods: a novel self-adapting semi-supervised clustering method and application case study

Article

Accepted Version

Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

Zhang, X., Ramirez-Mendiola, J. L. ORCID: https://orcid.org/0000-0001-7666-7440, Li, M. and Guo, L. (2022) Electricity consumption pattern analysis beyond traditional clustering methods: a novel self-adapting semi-supervised clustering method and application case study. Applied Energy, 308. 118335. ISSN 0306-2619 doi: https://doi.org/10.1016/j.apenergy.2021.118335 Available at https://centaur.reading.ac.uk/102324/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1016/j.apenergy.2021.118335

Publisher: Elsevier

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Electricity consumption pattern analysis beyond traditional clustering methods: A novel self-adapting semi-supervised clustering method and application case study

Xiaohai Zhang[1], José Luis Ramírez-Mendiola[2], Mingtao Li[1*], Liejin Guo[1]
[1] International Research Center for Renewable Energy (IRCRE), State Key Laboratory of Multiphase Flow in Power Engineering (MFPE), Xi’an Jiaotong University, Xi’an, Shaanxi 710049, China.
[2] School of the Built Environment, University of Reading, Whiteknights, Reading RG6 6AF, United Kingdom.

[*] Corresponding author.
E-mail address: mingtao@xjtu.edu.cn

Abstract

The fast-paced informatization of power systems across the world provides an unprecedented amount of data, which greatly facilitates their study and offers in turn the possibility to assist in the transition towards truly smart, low-carbon energy systems. In this context, the use of clustering methods for the study of household Electricity Consumption Behaviour (ECB) proves highly beneficial as it facilitates, among other things, more effective deployment of distributed renewable energy assets, development of differentiated tariff policies and load forecasting. However, the similarity metrics used in traditional clustering methods have difficulties in accurately capturing the time variability of electrical load profiles. In order to address this problem, we developed a novel semi-supervised automatic clustering method based on a self-adapting metric learning process. The proposed method is a bespoke application to the analysis of electricity demand load patterns that combines the recently developed Deep Linear Discriminant Analysis algorithm for supervised learning with the data-adaptive Affinity Propagation clustering algorithm (DLDA+AP), and achieves high-quality automatic clustering with an accuracy that is 75 percentage points higher than traditional methods such as k-means, on average. Based on this bespoke method, a unified load dictionary which captures the mainstream daily electricity consumption patterns of 5566 households in London was produced. Through the analysis of the load dictionary and household daily electricity consumption, it’s possible to build a complete ECB profile for the households in the sample dataset. Furthermore, combining the 206 household properties which were found to be strongly correlated with the ECB, this method provides a practical approach to residential customer segmentation for the electricity market.

Keywords: Household electricity consumption behaviour; Semi-supervised clustering; Metric learning

Nomenclature

ACC	Absolute value of correlation coefficient	KNN	K-Nearest Neighbors Classification
AMI	Adjusted mutual index	L2	Euclidean distance
AP	Affinity propagation	M	Mean of daily electricity consumption
ARI	Adjusted rand index	PNC	Peak number category
DBI	Davies-Bouldin index	P25	25 % percentile of daily electricity consumption
DEC	Daily electricity consumption	P50	50 % percentile of daily electricity consumption
DECP	Daily electricity consumption pattern	P75	75 % percentile of daily electricity consumption
DLDA	Deep linear discriminant analysis	SC	Silhouette coefficient
DNN	Deep neural network	SSD	Sum of squared distances of samples to their closest cluster center
DTW	Dynamic time warping		
ECB	Electricity consumption behaviour	Std	Standard deviation of daily electricity consumption
ECPk	Electricity consumption peak		

1. Introduction

The development of photovoltaic and wind power has brought green and low marginal cost electricity supply to the power system [1]. However, their inherently transient and fluctuating nature also poses a significant challenge to the stability of conventional power grids [2]. To address this issue, in addition to upgrading the grid hardware, a range of supply and demand matching technologies have been proposed, such as energy storage [3], demand response [4, 5], generation forecasting [6] and load forecasting [7, 8]. Coupled with the rise of "smart" devices, connected via the Internet of Things, along with artificial intelligence technologies that have provided researchers with a wealth of data and analytical tools, there has been an increased interest in researching potential benefits and applications of such technologies for demand-side management [9, 10].

When it comes to the diversity of electricity demand, the residential sector ranks second only after the industrial sector [1]. The vast market it represents and its inherent similarities make it one of the most promising targets for the implementation of demand-side management regulations. However, the success of such interventions vastly depends on how well understood the current energy consumption behaviours are, and their adaptability to any measures put in place. In this regard, a systematic exploration of the characteristics of household load profiles may provide the elements for a more helpful segmentation of the households that make up the residential electricity market. This in turn offers the possibility to identify specific groups (e.g. heavy users during peak periods) which can then be targeted for reduction measures.

In addition, a more in-depth analysis of the differences between the load profiles across households within the same group may reveal changes in electricity consumption behaviour (ECB) which can help reduce overall electricity consumption or improve efficiency through, for instance, changing the hours of use of specific equipment [11]. In summary, there are strong incentives for a more in-depth analysis of household ECB, as this forms the basis for smart tools such as energy storage strategies, modelling residential electricity demand and demand response [7, 12].

In recognition of this, several studies have used “traditional” clustering methods to explore the ECB of households [13]. However, most previous studies have focused on the classification of households based on their average load profile to represent the ECB [14, 15]. Such an approach may be suitable for describing the aggregated ECB, but the use of average profiles means that the variation within households is masked; while this approach provides a relatively straightforward classification of households’ ECB, it is clearly done at the expense of accuracy. The prevalence of such an approach is largely due to the complexity of the variability within and across load profiles, as well as the inability of traditional clustering similarity metrics to effectively capture the temporal variations between different patterns and thus accurately classify them into different categories [11, 16, 17].

To counter this issue, much effort has been devoted to comparing the effects of various time-series similarity metrics, such as Dynamic Time Warping (DTW), Pearson correlation coefficient, and other distance metrics such as the Euclidean distance (otherwise known as the L2 norm), in order to choose the best performing one for clustering purposes [18]. These traditional similarity metrics, however, may still be inadequate for the analysis of household electrical load profiles, which consequently reduces the effectiveness of clustering methods. Moreover, as with any other data-driven approach, clustering results are highly sensitive to the quality of the data used, and the previous studies have been characterised by the following two issues. Firstly, the sample sizes of the groups of households studied have usually ranged from tens to a few hundred households. Secondly, the duration of the monitoring periods has usually been less than a year. These two issues in combination mean that the overall amount of data used in said studies is rather small for clustering purposes, which poses serious limitations on the representativeness of the findings and, in some cases, even leads to significant group bias; this may well be the main reason for conflicting findings across a number of studies [19].

This paper presents a unique methodological contribution to the energy research literature, in which:

- a novel semi-supervised automatic clustering method based on a self-adapting metric learning process has been developed and tailored to the analysis of household load profiles;
- the proposed method has been applied to the study of ECB in London as a case study; and
- the results of its application have been compared to traditional clustering methods to assess its performance.

The proposed method, henceforth referred to as DLDA+AP, combines the recently developed Deep Linear Discriminant Analysis (DLDA) algorithm for supervised learning [20] with the data-adaptive

Affinity Propagation (AP) clustering algorithm [21], and demonstrably achieves superior classification results when compared to traditional approaches to clustering-based ECB analysis.

The method proposed in this paper is a response to the challenges imposed by ill-suited similarity metrics used in previous clustering studies of household ECB. Specifically, DLDA and AP work in tandem; DLDA is used to tailor a similarity metric best suited to the analysed data aided by a complimentary labeled daily load sample dataset; this tailored similarity metric is then used by AP to generate a load dictionary representing mainstream daily electricity consumption patterns (DECPs) based on the original dataset. The resulting load dictionary then provides the basis for an exploratory analysis of the characteristics and commonalities of ECB across households, as well as their potential correlation with household properties. Moreover, it allows for a more in-depth analysis of the composition of daily electricity consumption (DEC) and DECPs which capture the hourly variation of electricity consumption throughout the day.

The specific case study serves as an illustration of the advantages of the use of DLDA+AP for the study of residential electricity consumption patterns, and allows us to place this novel method in the context of the traditional clustering approaches, such as k-means-based clustering, by means of a detailed comparison of their performance in the classification of household ECB based on the same dataset. In doing so, this paper lays the groundwork for the adaptation of a method expressly designed for the analysis of electricity consumption patterns to other sectors, as well as other regions and countries.

What remains of this paper is arranged as follows: Section 2 provides a brief review of related work; Section 3 introduces the data used for the study and pre-processing methodology, as well as the elements of the semi-supervised clustering model and the classification model; Section 4 evaluates the performance of the trained models; Section 5 analyses the clustering results and household ECB; finally, Section 6 offers our concluding remarks.

2. A brief review of clustering-based analyses of household ECB

Clustering methods have found widespread applications in a number of fields, and the analysis of energy demand patterns is no exception. In what follows, we review some of the most recent studies on electrical load profile clustering, and we focus on three key aspects of clustering: data features, similarity metrics and clustering algorithms.

When it comes to household ECB data, the defining features are the electricity consumption over a fixed length of time or the percentage of electricity consumption corresponding to a given period, such as daily, weekly, seasonal and annual samples, with temporal resolutions ranging from 15 minutes to halfhour or an hour [14]. Such features mainly reflect differences in the peaks and troughs of electricity consumption and can be understood intuitively as differences in the shape of the load curve. Markovič et al. [22] used the average weekly profile and the annual profile to explore the relationship between short-term and long-term electricity consumption patterns and thus improved the clustering of household customers. Other methods such as dimension reduction and feature extraction, known as alternative time-series representation have been used as well [18]. Satre-Meloy et al. [23] proposed a clustering method for cumulative electricity consumption over time as a feature, which had a clear physical meaning and achieved promising results.

In terms of similarity metrics, the Euclidean distance (L2), Dynamic Time Warping (DTW) and Pearson correlation coefficient are the most frequently used [18]. Iglesias et al. [16] compared L2, DTW, Pearson correlation coefficient and Mahalanobis distance as similarity metrics for clustering electrical load samples; the results showed that L2 outperformed the others, while DTW was also effective in some cases. However, given that the computational complexity of DTW is much higher than that of L2, most studies use L2 as their similarity metric. Nevertheless, the way in which L2 is calculated limits its ability to accurately describe the differences in temporal variation across electricity consumption patterns. This is particularly the case where the load samples have a large and relatively concentrated proportion of electricity consumption peaks (ECPks), as L2 cannot effectively reflect the temporal differences in where the peaks are located [23].

The algorithms used for household ECB clustering can be broadly classified into four categories: Partitioning, Hierarchical, Model-based and Density-based clustering [18]. Partitioning clustering is typically represented by the K-means algorithm which is usually based on L2 and is susceptible to outlier interference. More importantly, this kind of algorithm usually requires the number of clusters to be specified in advance. Hierarchical clustering can be performed in two ways: bottom-up and top-down. The former groups highly similar elements one by one, while the latter discards the elements with the lowest similarity one by one. The method is strongly influenced by changes in the samples. However, an arbitrary number of clusters can be chosen as required, and thus it is often used for similar category merging. Kwac et al. [24] first used adaptive K-means with constraints to cluster a large number of daily load samples to obtain over 100,000 highly similar classes, and then used bottom-up hierarchical clustering to obtain 1000 merged classes that satisfied its set constraint error of 5% to generate a load dictionary. The most commonly used Model-based approach for household ECB clustering is self-organising maps (SOM), which is based on neural networks that update the network weights to form suitable clustering centers by comparing the distance between the input samples and the output vectors. The method is based on assumptions about the topology of the data, and its accuracy is usually similar to that of K-means. McLoughlin et al. [25] evaluated the clustering effectiveness of K-means, K-medoids and SOM using the Davies-Bouldin index and, based on this, decided to cluster daily electricity load samples of Irish households using SOM, obtaining 10 typical classes. Density-based clustering, such as DBSCAN, is often used to find anomalous samples [26]. However, due to the high computational complexity of such methods, they are rarely used in clustering household ECB with large sample sizes [18]. In those studies, it was found that the most commonly used methods for clustering household ECB profiles are still classical clustering algorithms, which might be, at least in part, due to the slow development of alternative clustering algorithms [18]. Most algorithms perform large amounts of repetitive clustering by changing the parameters of the algorithm and thus, based on evaluation indices and expert knowledge, allow to choose a suitable number of clusters manually. This approach, however, creates a clear obstacle to the development and application of data-adaptive clustering.

Some studies have shown that peak features can effectively describe the differences between electricity consumption patterns [14, 24]. However, traditional similarity metrics, including the high-performing L2, perform poorly in describing the temporal variation between peak features. To address this problem, we first label the training dataset with the value and time period of the ECPks of daily load samples to guide the supervised learning algorithm DLDA to learn the most suitable similarity metric. Then, the data-adaptive AP clustering algorithm is used to perform clustering without manually specifying the number of clusters. This process allows for fully automated model training and clustering on different datasets.

An in-depth description of the development of the DLDA and AP algorithms falls beyond the scope of this paper, but can be found elsewhere [20, 21, 27]. Likewise, for further details about clustering-based research in the context of energy consumption analysis, the reader is pointed to the following references: [7, 13, 14, 18, 28].

3. Data and Methods

3.1 Description and processing of data

3.1.1 Description of data

The data used for the analysis presented in this paper stems from the Low Carbon London project, a study led by the UK Power Networks (one of the UK's Distribution Network Operators), during which the electricity consumption of 5,566 London households was monitored between November 2011 and February 2014 [29]. Households were recruited such that they comprised a balanced sample, representative of the Greater London population [4], and grouped based on the Acorn household classification methodology [30]. The dataset contains a vast array of electricity consumption load profiles (half-hourly energy consumption in kWh), marked with a unique household identifier, date, tariff type [31, 32] and the corresponding Acorn group.

The Acorn classification methodology segments consumer households into 17 groups from ACORN-A to ACORN-Q. By analysing significant social factors and population behaviour, it provides precise information and an in-depth understanding of the different types of people living in London. The classification is based on 15 major items and 84 subitems for a total of 826 options as shown in Table A-1 in the Appendix. The major items include Population (P), Housing (Hou), Family (Fam), Economy (Eco), Education (Edu), Health (Hea), Transport (T), Marketing Channels (MC), Finance (Fin), Digital (D), Shopping (S), Contact (Con), Environment (Env), Community Safety (CS) and Leisure Time (LT). The value of each option is a relative value against the national average. Generally speaking, there is a decreasing trend of wealth from ACORN-A to ACORN-Q.

3.1.2 Preprocessing of data

Firstly, the half-hourly energy consumption records were turned into hourly profiles and then split into 24-dimensional load samples corresponding to the different days of the monitoring period. The hourly load samples were then normalised to create the corresponding DECP of individual households, in order to examine the variation in hourly electricity consumption throughout the day. Thus, for the i-th load sample, the d_i denotes the total DEC, $h_i(t)$ denotes the electricity consumption at time t (in hours), and $p_i(t)$ denotes the electricity consumption at time t (in hours) as a percentage of the total DEC. These relations are expressed by the following equations:

$$d_i = \sum_{t=1}^{24} h_i\,(t) \quad and \quad p_i(t) = \frac{h_i(t)}{d_i} \times 100\% \tag{1}$$

When analysing the dataset, we excluded samples based on the following criteria:

- household samples with load profiles for less than a week;
- almost zero d_i load samples with DEC lower than 0.024 kWh. The threshold represents 1 W per hour for standby electrical appliances like smart meter.

Besides, we found that group ACORN-U lacked the relevant information. Therefore, this group was only used for the analysis of properties not related to Acorn classifications.

By clustering the normalised load samples, we can obtain a load dictionary that represents the household DECP for the whole London area. Based on the unified load dictionary patterns, we are able to analyse the DECP composition of each household. Further, in addition to the DEC, a complete ECB profile of the household can be created.

After cleaning and preprocessing the raw data, we were left with a dataset containing the electricity consumption records of 5,557 households across 18 Acorn groups, which contains nearly 3.5 million daily load profile samples. Figure 1 shows that the household sample distribution is consistent with the load sample distribution after cleaning of the data, ensuring the dataset is representative of the London population.

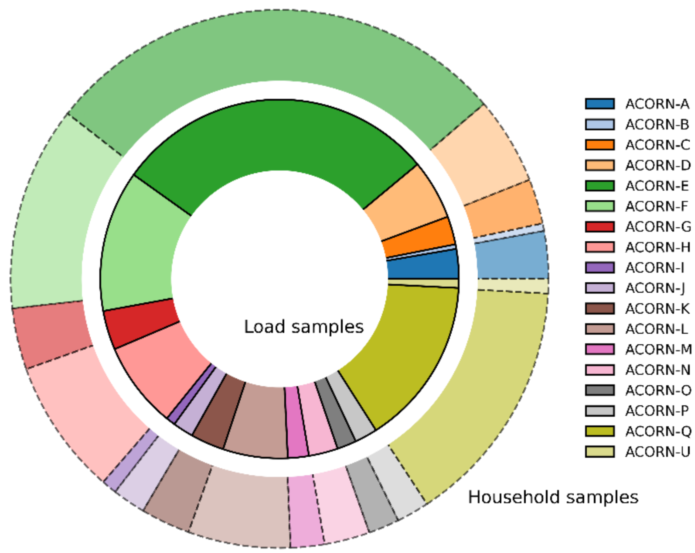


Figure 1 - Distribution of household samples and load samples across Acorn groups.

3.1.3 Building labeled dataset

To provide DLDA with labeled load samples for metric learning, a labeling method for the value and time period division of the ECPk features is proposed. The proposed labeling method essentially consists in comparing individual load sample plots to a reference grid which divides the plotting canvas into the regions that capture the most (and least) interesting features of the plotted load samples.

The identification of these regions of interest is based on empirical experience and partial visualisation of a subset of samples. Along the vertical axis, we divided the potential values of ECPk into 3 bands, and along the horizontal axis, we divided the 24 hours in a day into 8 time periods as detailed in Table 1 and Table 2; see also Figure 2 for a graphical representation. It is worth noting that the choice of the boundaries of the regions of interest defined above is in fact arbitrary, and that as long as the sub-divisions can reliably distinguish between different feature categories (e.g. peaks), any other labeling grid can be used to guide training of the semi-supervised clustering model. However, since the performance of the trained model greatly depends on the labeled dataset used, it is advisable to try and make the best possible choice.

As discussed above, in the case of the proposed labeling method, the regions of interest were determined purely based on the characteristics of the empirical data used in this study, supported by our knowledge of load pattern dynamics. It should also be noted that for the purposes of our analysis, the ECPk is defined as the time period's maximum hourly electricity consumption $h_i(t)$ higher than 6.255% of the DEC (d_i), in which the threshold value represents 1.5 times the average percentage of the 24-hour period. In addition, in order to avoid the peaks appearing near the dividing line of adjacent intervals along the vertical axis (which leads to blurred peak features), we narrowed down the value intervals for labeling the peak, which were called label range, as shown by the pink stripes in Figure 2; only the sample with all peaks located in the label range of 3 levels or without peaks (e.g., all the value lie in average interval) would be selected and labeled. Figure 3 provides two examples of the application of these labeling criteria to produce two labeled samples.

Table 1 - Value division of ECPk.

Value level	Value range (%)	Label range (%)	Notes
Average	2.085 - 6.255	Same	$100/24 \pm 50\%$
Level-1	6.255 - 12.51	8.255 - 10.51	$12.51 = 6.255 \times 2$
Level-2	12.51 - 50.00	14.51 - 40.00	$50 \approx 100 - 2.085 \times 23$ hours
Level-3	50.00 - 100.00	Same	

Table 2 - Time period division of the day.

Time period	t-1	t-2	t-3	t-4
Hour	1-3	4-7	8-10	11-13
Time period	t-5	t-6	t-7	t-8
Hour	14-16	17-19	20-22	23-24

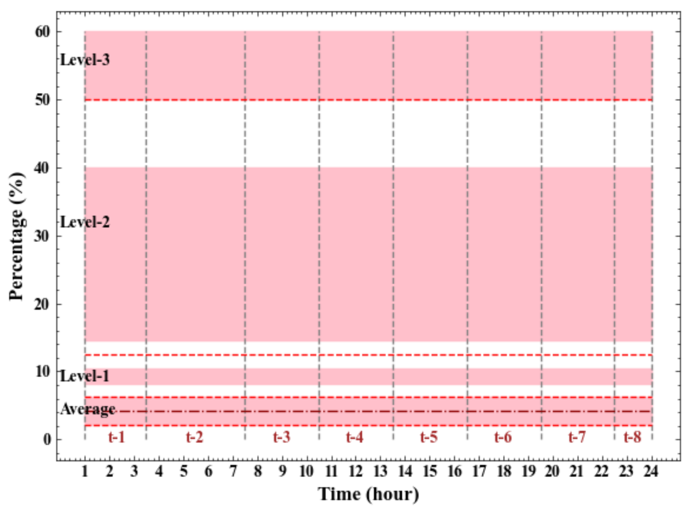
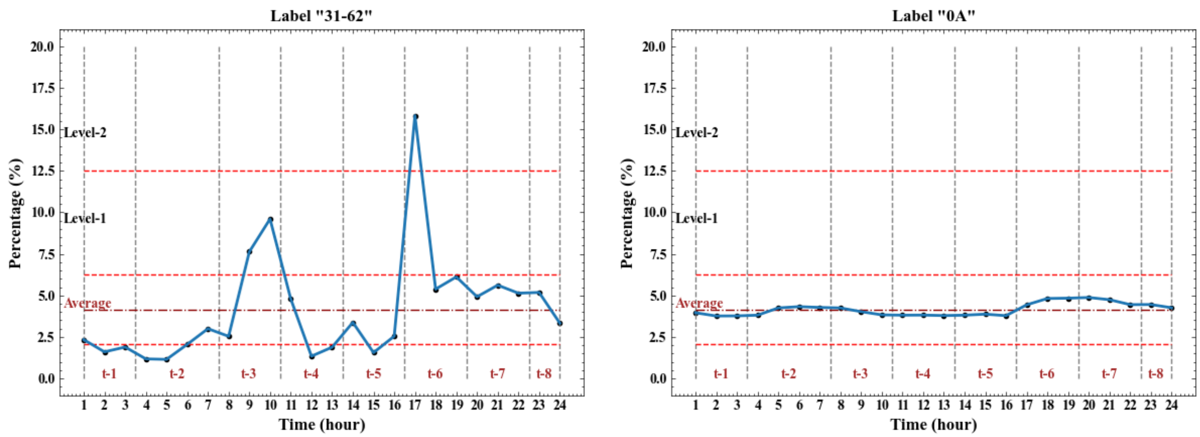


Figure 2 - Visualization of labeling principles.



(a) Label "31-62" (b) Label "0A"

Figure 3 - Examples of labeled samples.

Figure 3 (a) label "31-62" indicates the sample has two ECPks, one at level-1 in the value range of time period t-3 and one at level-2 in the value range of time period t-6. Figure 3 (b) label "0A" indicates the sample has no ECPks.

After labeling the load samples, we obtained 1979 classes containing 568,165 labeled samples, accounting for 16.45% of the total sample. To further filter out samples representing the mainstream DECPs, we selected classes with sample size greater than 365, and thus obtained 160 classes with a total of 514,595 labeled samples, representing 90.57% of the total labeled samples. Considering the unbalancing of class size, we conducted random downsampling to obtain a balanced dataset containing 160 classes with 367 samples in each class which accounts for 11.41% of mainstream labeled class samples. Figure 4 shows the differences in the distribution of mainstream labeled samples represented by peak numbers before and after balancing. The Kolmogorov-Smirnov (KS) test [33] between each downsampled and unsampled class calculated a minimum p-value of 0.66 (as shown in Figure 5), indicating a consistent distribution between the two, which also demonstrates the validity of ECPk as a representation of the sample's features. Finally, a balanced labeled dataset containing 58,720 samples in 160 classes was built for the training and testing of the semi-supervised model.

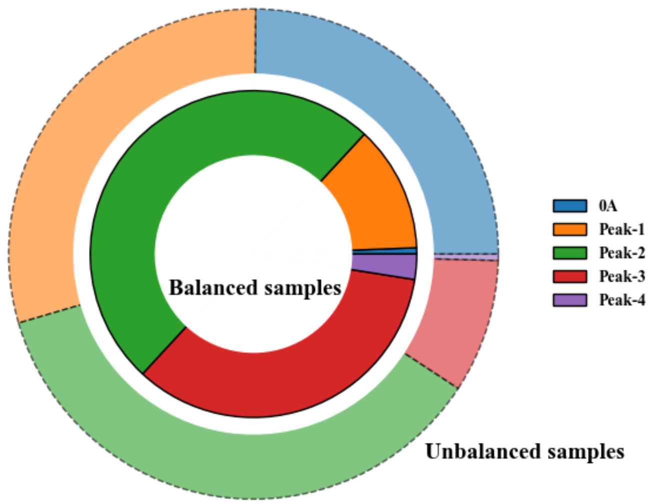


Figure 4 - Differences in the distribution of mainstream labeled samples represented by peak numbers before and after balancing.

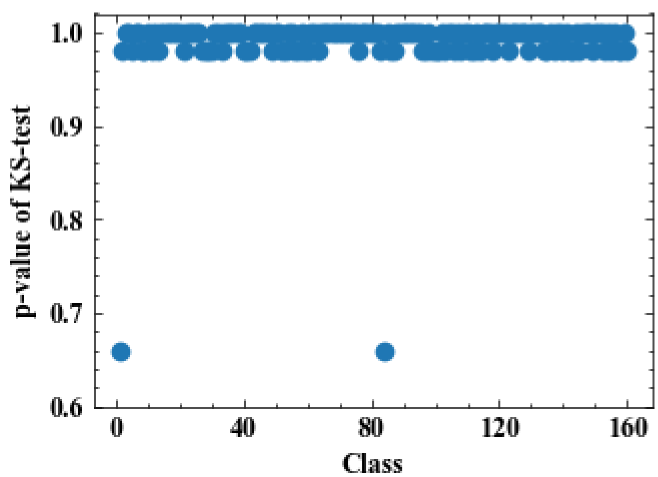


Figure 5 - *p*-values of KS-test between unsampled and sampled data in each mainstream labeled class.

3.2 Semi-supervised clustering method

Figure 6 provides a graphical representation of the whole workflow - from data preprocessing to final segmentation of households – that was followed in the development. The proposed semi-supervised clustering method combines DLDA with AP, and is trained using the balanced labeled dataset. The trained model is further evaluated by means of comparing it with other clustering methods on new datasets (e.g., mini and mixed datasets for blind validation). During the supervised learning stage, a DLDA model is trained to find an effective projection space that can enhance the similarities of intra-class samples and discrepancies of inter-class samples, which is also considered to be the metric learning process here as the new distance between two samples is calculated by using the non-linear transformation of DLDA followed by L2. Both the DECP load samples and distance metric are then passed on to the next stage, where AP is used to carry out the unsupervised clustering. Since AP is a data-adaptive clustering algorithm, it does not require the number of clusters to be specified in advance and provides stable clustering results. This enables an automatic evaluation of DLDA during the training process and thus helps to achieve an adequate trade-off between fitting and generalisation. Therefore, a well-trained semi-supervised clustering model can adapt to DECP load samples in practical clustering applications.

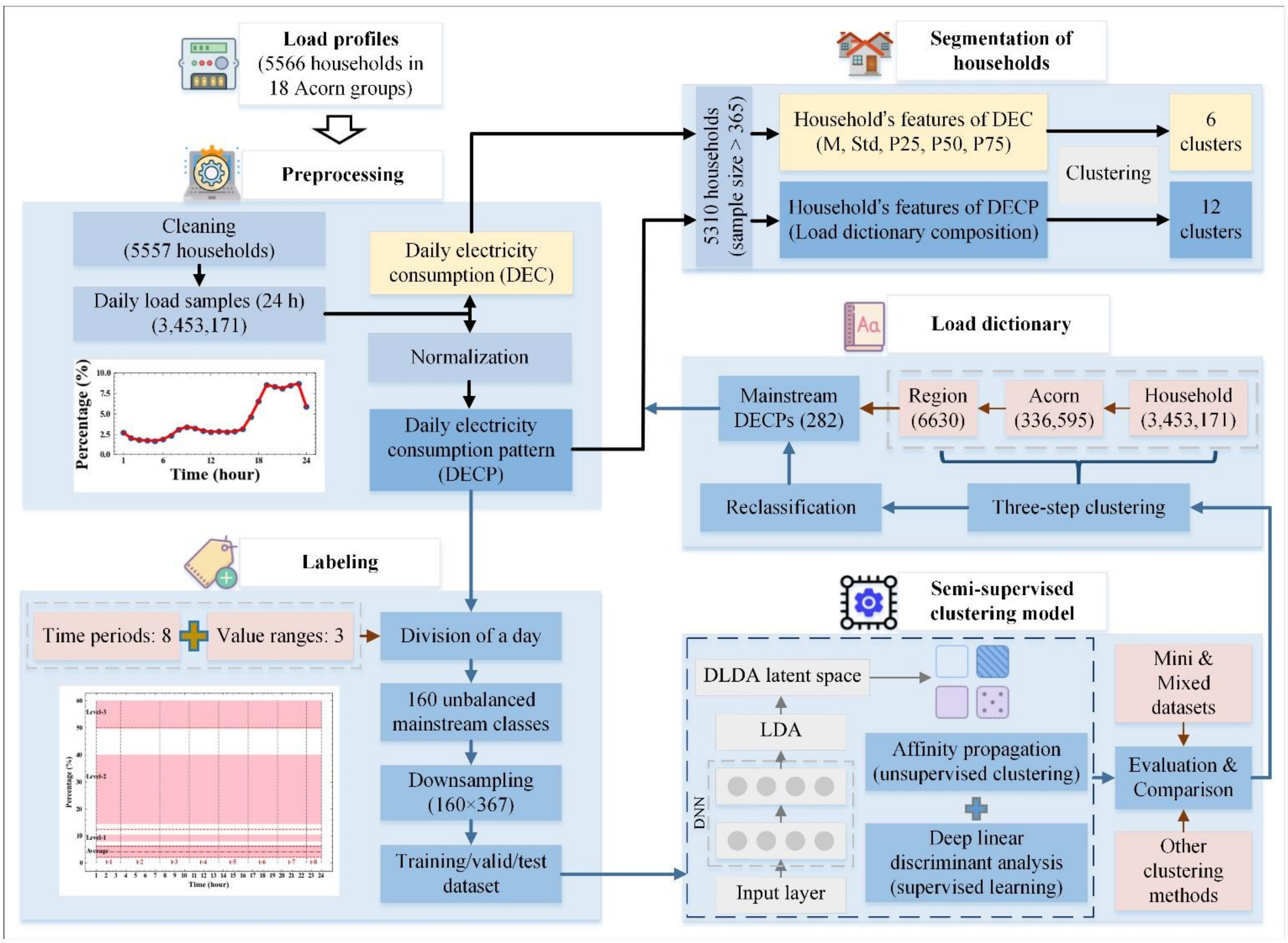


Figure 6 - Workflow of proposed clustering strategy.

3.2.1 Deep linear discriminant analysis (DLDA)

Classic Linear Discriminant Analysis (LDA) was originally proposed by Fisher [34] in 1936. The aim of LDA is to find a linear projection that maximizes inter-class scatter and minimizes intra-class scatter by maximizing the Fisher number. LDA is able to find the optimal decision boundaries when the data in different classes have the same prior distribution and possess Gaussian distributions with the same covariance. DLDA, proposed by Dorfer et al. [20], is a nonlinear extension of classic LDA, which takes the eigenvalue solutions of LDA to build the objective function for a deep neural network (DNN) , as shown in Figure 7. In this work, the full DLDA model is generated by adding a LDA transformation (LDA layer) to the output of DNN, which makes full use of the nonlinear transformation ability of DNN to help classic LDA find a better projection space with higher inter-class scatter and lower intra-class scatter.

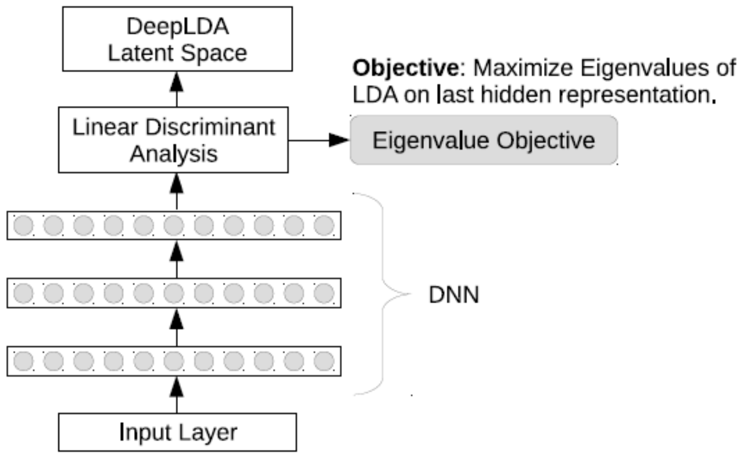


Figure 7 - Schematic sketch of the DLDA algorithm [20].

In analytical terms, the DLDA algorithm can be summarised as follows.

Let $x_1, \dots, x_n = X \in \mathbb{R}^{n \times d}$ denote a set of N samples belonging to C different classes $c \in \{1, \dots, C\}$.

With a linear projection matrix $W \in \mathbb{R}^{d \times (C-1)}$, the elements of X are transformed into the projection space. The Fisher number $J_F(W)$ is defined as:

$$J_F(W) = \frac{S_b{}'}{S_w{}'} = \frac{W^T S_b W}{W^T S_w W} \qquad (2)$$

where $J_F(W)$ is the ratio of between-scatter matrix $S_b{}'$ and within-scatter matrix $S_w{}'$ in the projection space. Using the linear transformation W, $S_b{}'$ and $S_w{}'$ can be calculated from the corresponding between-scatter matrix S_b and within-scatter matrix S_w in the original space where the elements of X are located.

Let $\bar{X}$ and $\bar{X_c}$ be the mean-centered observations of the entire set X and class c's elements, X_c, respectively. S_b and S_w can be calculated as follows.

$$S_c = \frac{1}{N_c - 1} \bar{X}_c^T \bar{X}_c \qquad (3)$$

$$S_w = \frac{1}{C} \sum_c S_c \qquad (4)$$

$$S_t = \frac{1}{N-1} \bar{X}^T \bar{X} \qquad (5)$$

$$S_b = S_t - S_w \qquad (6)$$

The objective function of LDA is to maximize $J_F(W)$.

$$\underset{W}{argmax} J_F(W) = \underset{W}{argmax} \frac{W^T S_b W}{W^T S_w W} \quad (7)$$

The solution of $\underset{W}{argmax} J_F(W)$ can be transformed to the general eigenvalue problem $S_b e = \lambda S_w e$. The linear projection matrix W is the set of eigenvectors e associated with this problem.

For DLDA, in order to update the gradient of parameters in the DNN using back-propagation in an end-to-end fashion, a new loss function was proposed (Equation (8)).

$$l(H) = \frac{1}{k}\sum_{i=1}^{k} \lambda_i \quad with \ \{\lambda_1, \dots, \lambda_k\} = \{\lambda_j | \lambda_j < min\{\lambda_1, \dots, \lambda_{C-1}\} + \varepsilon\} \quad (8)$$

Where λ is the eigenvalue of $S_b e_i = \lambda_i (S_w + \alpha I) e$ based on the output $H \in \mathbb{R}^{n \times h}$ from the top hidden layer in a DNN.

Adding a multiple of the identity matrix αI to the within-scatter matrix S_w could enhance calculation stability and optimize for small eigenvalues [27]. The objective function $\underset{\Theta}{argmax}(l(H))$ focuses on the optimization of smallest k eigenvalues, which do not exceed a certain threshold ε, of all $C - 1$ available eigenvalues.

During the training process of DLDA, the AP is used to achieve automatic and stable clustering of the final output from the LDA layer. Therefore, the validation of the model depends on the accuracy of the clustering results. If the results show high consistency with the true assignments, we assume the training process has been completed.

3.2.2 Affinity propagation (AP)

AP is a clustering algorithm that takes similarity between pairs of data points as input and clusters data points by passing and updating messages between them until high quality clusters emerge. Therefore, it does not need to prespecify the number of clusters and define an initial set of exemplar elements as with common k-centers clustering methods. It has been proved effective in detecting genes in microarray data, clustering images of faces, and identifying cities that are efficiently accessed by airline travel, and shows much lower error and time cost than other methods [21].

In analytical terms, the AP algorithm can be summarised as follows.

Let $x_1, \dots, x_n = X \in \mathbb{R}^{n \times d}$ denote a set of unlabeled samples. For the objective function of minimizing the sum of squared error (L2) from each data point to its nearest exemplar, the similarity for x_i and x_k is defined as $s(i,k) = -||x_i - x_k||^2$. And for the similarity of point x_k itself, $s(k,k)$ is defined as a real number called "preference", which determines the likelihood of x_k becoming an exemplar and will influence the final number of clusters. The shared value of this preference is the median of input similarities resulting in a moderate number of clusters or the minimum of input similarities resulting in a small number of clusters as usual.

The message passing between data points contains two parts which reflect different kinds of competition respectively. One is the "responsibility", $r(i,k)$, which captures the information offered to x_k by x_i; it determines the probability of x_k to serve as exemplar for x_i and takes into account other potential exemplars for x_i. The other one is the "availability", $a(i,k)$, passed from x_k to x_i, which shows the probability of x_i to choose x_k as its exemplar; it takes into account the situation in which other points choose x_k to be their exemplar. Figure 8 shows the diagram of sending responsibilities and availabilities. When the passing and updating iteration converges, and a solution point x_k of $\underset{k}{argmax}(r(i,k) + a(i,k))$ is found, it means that point x_k should be x_i's exemplar. And if $i = k$, it means x_i should be an exemplar.

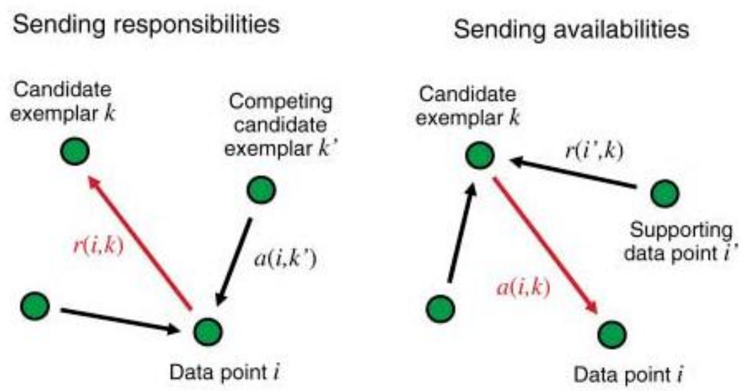


Figure 8 - Diagram of message passing [21].

For updating responsibility $R \in \mathbb{R}^{n\times n}$ and availability $A \in \mathbb{R}^{n\times n}$, the algorithm needs to compute the similarity between data points and set preference to offer the input similarity matrix $S \in \mathbb{R}^{n\times n}$ in the beginning; A is initialized as zero matrix. Firstly, R is updated using Equation (9).

$$r_{t+1}(i,k) = \begin{cases} s(i,k) - \max\limits_{k'\neq k}\{a_t(i,k') + s(i,k')\}, & i \neq k \\ s(k,k) - \max\limits_{i'\neq k} s(i',k), & i = k \end{cases} \tag{9}$$

Then, A is updated using Equation (10).

$$a_{t+1}(i,k) = \begin{cases} min\{0, r_t(k,k) + \sum_{i'\notin\{i,k\}} \max\{0, r_t(i',k)\}\}, & i \neq k \\ \sum_{i'\neq k} max\{0, r_t(i',k)\}, & i = k \end{cases} \tag{10}$$

To avoid numerical oscillations, a damping factor $\beta \in (0,1)$ is introduced. Thus, the new R's and A's are calculated as follows:

$$r_{t+1} = (1-\beta) r_{t+1} + \beta r_t \tag{11}$$

$$a_{t+1} = (1-\beta) a_{t+1} + \beta a_t \tag{12}$$

The algorithm stops when any of the following conditions is met: (a) a fixed number of iterations is reached; (b) changes of message are lower than a threshold; (c) exemplars do not change within a fixed number of iterations.

In what follows, DLDA transformation followed by L2 between load samples was taken as the similarity measure, and the median of input similarities was set as the preference of all points to achieve a trade-off between quality and number of clusters.

3.3 K-Nearest Neighbors Classification (KNN)

KNN is a kind of instance-based learning or non-generalizing learning model, which means that it does not try to construct a general internal model, but simply stores instances from training samples [35]. The classification procedure is done through "voting" from k nearest instances around the sample. If the voting is not weighted, the predicted sample would be classified to the majority class of k nearest neighbors. If voting is weighted, the vote of the closer neighbor would normally have a heavier weight.

KNN was used here to reassign load samples into the formed load dictionary classes to improve the quality of mainstream clusters. For the purposes of implementing KNN, the DLDA transformation followed by L2 was taken as the distance metric, and the value of k was set to 1, which means each data sample would be assigned to its nearest load dictionary class.

4. Evaluation of clustering and classification

4.1 Evaluation criteria

External clustering criteria, namely Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI), were used to evaluate the performance of the semi-supervised model and classification model since these models use samples with knowledge of the ground truth classes. For clustering samples without label, internal clustering criteria, namely the Silhouette Coefficient (SC) and Davies-Bouldin Index (DBI), were used to evaluate the quality of clusterings. A brief description of them has been summarized in Table A - 2 in the Appendix. Bigger values of ARI, AMI, and SC closer to 1 and smaller value of DBI closer to 0 mean better results.

4.2 Training and testing of semi-supervised model

All the models were implemented in Python (v3.7), using the packages scikit-learn (v0.23) [35] and tensorflow (v2.2.0) [36] on a machine with a NVIDIA K80 GPU.

4.2.1 Optimization of hyperparameters in DLDA

As part of the DNN of the DLDA model, we built a "dense layer - batch normalization layer - activation layer" block to enhance the numerical stability. In what follows, the number of layers refers to the number of this block. To ensure better model training, we used the Python package hyperopt [37] which provides methods based on Bayesian optimization to find optimal candidates of hyperparameters (as listed in Table 3) and allows for making improvements based on them manually. The objective function of optimization is to maximize the mean of ARI and AMI between true assignments of samples and AP clustering results based on the output of DLDA under 3-fold cross validation (CV). The activation function was set as sigmoid for the top hidden layer and as ReLU for the remaining layers of the DNN. The Adam optimization algorithm was used as optimizer. The labeled dataset was split, using three-quarters of it as the training set and one-quarter as the test set, which was taken as the input for the model without further processing.

Table 3 - Value range and optimized results of hyperparameters.

Parameter	Value range	Optimized result
Number of layers	[2,10] (Integer)	3
Units	[48,2400] (Integer)	2048
Epochs	[5,3000] (Integer)	950
Output dimensions of top hidden layer	[24,1200] (Integer)	1024
Regularization factor (l2)	$\{0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$	0
k value of Equation (8) (ϵ is not considered here)	[1,159] (Integer)	120

4.2.2 Evaluation of semi-supervised model

Using the optimized hyperparameters, we obtained the trained semi-supervised clustering model. The results in Table 4 demonstrate the model exhibits excellent generalisation on the test set with an ARI and AMI mean of 0.977, which is also better than its performance in 3-fold cross-validation. This indicates that the trained semi-supervised clustering model can effectively capture the peak features of load samples and thus differentiate them well, which essentially achieves the intended goal.

Table 4 - Evaluation of trained model in 3-fold cross validation and on the test set.

Criteria	CV-1	CV-2	CV-3	Test set
ARI	0.933	0.928	0.937	0.967
AMI	0.953	0.942	0.947	0.987
Average	0.943	0.935	0.942	0.977
Number of clusters (real number is 160)	218	223	219	181

To demonstrate the superior performance of the trained model, we compared it with other clustering methods including K-means using L2 and DTW distances respectively, AP, classic LDA combined with AP, only using the DNN part of DLDA combined with AP, as well as full DLDA combined with K-means. For comparison purposes, the series of steps associated with each of the discussed clustering methods is shown in Figure 9. The preference of AP was set as the median of input similarities. A mini labeled dataset for this blind validation was created by randomly selecting 10 load samples per class from the 160-class unbalanced labeled dataset described in the previous section. The selections of mini dataset and clusterings were repeated 10 times to ensure that the results were statistically sound. The results of these comparisons, which are summarized in Table 5 and Table 6, reveal that the trained semi-supervised clustering model (DLDA + AP) significantly outperforms other clustering methods and provides clustering results almost identical to the true assignments. Furthermore, another blind validation dataset called mixed dataset which contains a portion of the mini labeled dataset (14.9 %) and randomly selected unlabeled samples (85.1%) was built to evaluate the performance of the trained model in practical clustering, which was carried out 10 times independently as well. The results of this final test prove how accurate the clustering of labeled samples is, where a mean value of 0.964 for ARI and AMI was observed (See Table 7). Therefore, it can be concluded that the semi-supervised clustering model successfully finds a projection space which effectively separates samples between classes based on the prior knowledge provided by the labeled dataset, and achieves high-quality automatic clustering.

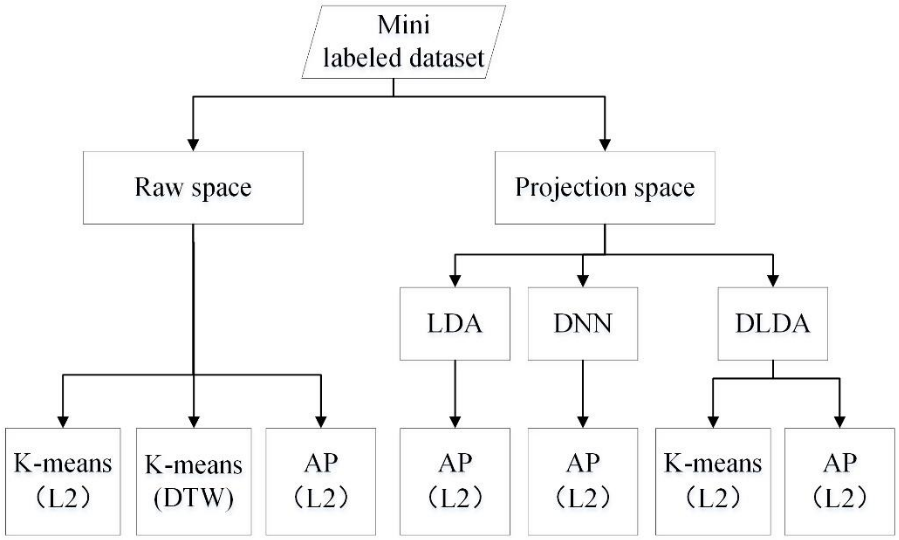


Figure 9 – Different key steps followed by different clustering methods used for comparison purposes.

Table 5 - Results of different clustering methods on mini labeled dataset (mean of 10 repeats).

Methods	Mean of ARI and AMI	Mean number of clusters
K-means (L2)	0.262	160 (Prespecified)
	0.147	20 (Unspecified)
K-means (DTW)	0.213	160 (Prespecified)
	0.158	10 (Unspecified)
AP (L2)	0.249	116
LDA + AP (L2)	0.264	113
DNN + AP (L2)	0.526	170
DLDA + K-means (L2)	0.988	160 (Prespecified)
	0.980	158 (Unspecified)
DLDA + AP (L2)	0.996	159

Notes: "Unspecified" in K-means methods means the best number of clusters is chosen according to the scores of SC, DBI and SSD (elbow principle) with number of clusters ranging from 2 to 200.

Table 6 - SC and DBI of AP and K-means clustering results based on DLDA on mini labeled dataset (mean of 10 repeats).

Criteria	SC	DBI
DLDA + K-means (L2) (Prespecified)	0.601	0.616
DLDA + K-means (L2)	0.604	0.613

(Unspecified)		
DLDA + AP (L2)	0.612	0.554
Best one	DLDA + AP	DLDA + AP

Table 7 - Evaluation of trained semi-supervised model on mixed dataset (mean of 10 repeats).

Criteria	SC	DBI	Mean of ARI and AMI (only labeled samples)
DLDA + AP (L2)	0.169	1.697	0.964

4.3 Three-step clustering

As the whole dataset contains over 3 million load samples, we proposed a three-step clustering strategy based on the trained semi-supervised clustering model to produce a load dictionary representing mainstream DECPs; the three steps correspond to household, Acorn group and region levels, respectively. Firstly, the load samples were clustered to obtain each household's cluster centers (average value) which were then used as the samples for clustering at the Acorn group level. Then the samples at the Acorn group level were clustered to obtain the cluster centers of each Acorn group which were then used as the samples constituting the region level. The final clustering results were obtained by clustering the samples at region level. The number of samples at household, Acorn group and region levels, as well as the final number of clusters are given in Table 8.

Table 8 - Number of samples at household, Acorn group and region level, and final clusters.

Level	Household	Acorn	Region	Final clusters
Size	3,453,171	336,595	6630	458

The clustering results at each level were evaluated and compared with the clustering results from the mini labeled dataset and mixed dataset as shown in Figure 10. The results show the clustering quality of the whole dataset is lower than that of the mini labeled dataset, but it's better than that of the mixed dataset. This is due to the fact that the labeled dataset contains only well-characterised samples of the dominant classes, while both the whole dataset and mixed dataset contain a large number of "noisy" samples. Although the physical meaning of ECB ensures the peak features of samples remain distinct, the noisy and outlier samples still affect the quality of clustering and reduce the scores of evaluation criteria, which is clearly reflected by the large number of clusters containing only a very small number of samples generated during the clustering process.

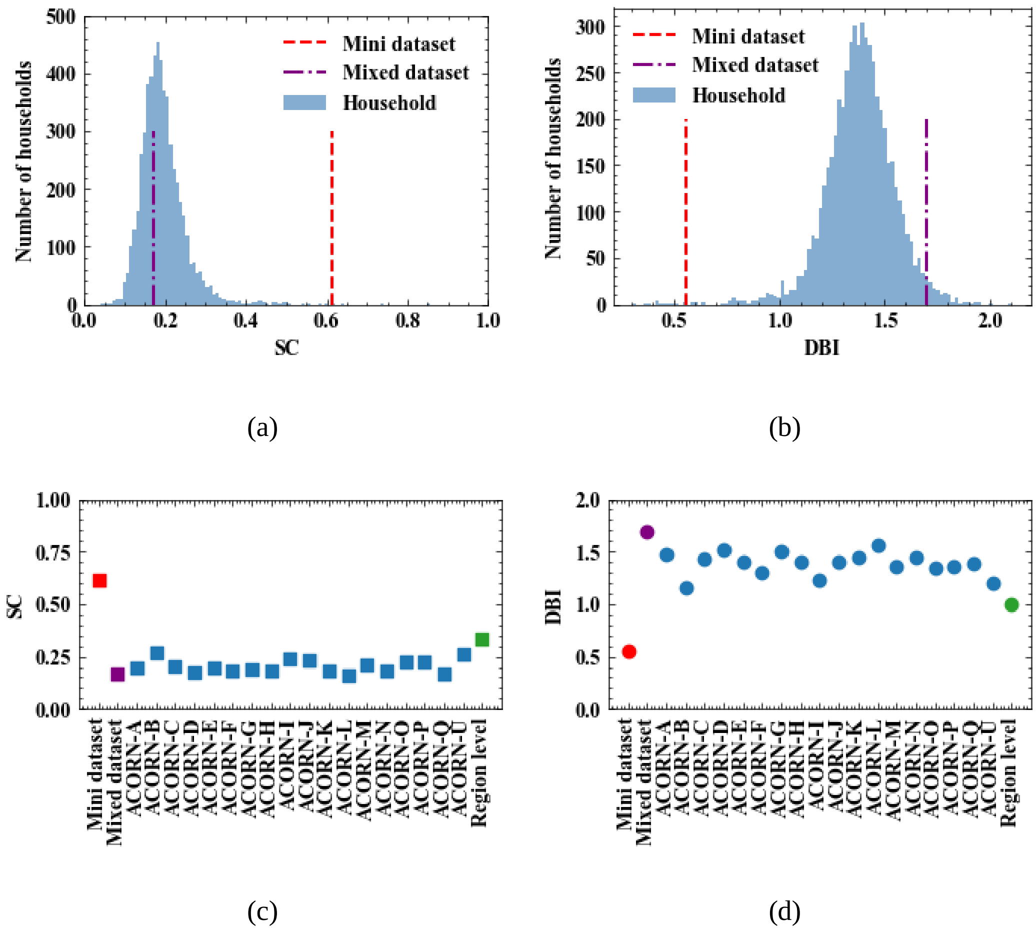


Figure 10 - Silhouette Coefficient (SC) and Davies-Boulding Index (DBI) for the three-step clustering results: SC (a) and DBI (b) at household level; SC (c) and DBI (d) at Acorn group and region levels.

4.4 Reclassification of samples

The total number of clusters at the region level (458) is much less than that of the originally labeled data (1979), but is almost 3 times that of the selected labeled dataset (160). This could potentially be attributed to either many outliers or low-frequency patterns incorrectly clustered together with mainstream pattern samples. Therefore, we make a reassignment of all samples to improve the quality of clusters.

As mentioned before, KNN with $k = 1$ and DLDA transformation followed by L2 was used to build the classification model. The training set was the cluster centers of final clusters. For testing the accuracy of the KNN model, samples and clustering results at region level were used as a rough test set. Results in Table 9 indicate most samples can be correctly classified. Considering the deficiency of test set (e.g. presence of outliers), we consider the accuracy of KNN is acceptable.

Table 9 - Evaluation of KNN model on test set.

Criteria	ARI	AMI	Accuracy

KNN	0.897	0.923	0.926

5. Results and Discussion

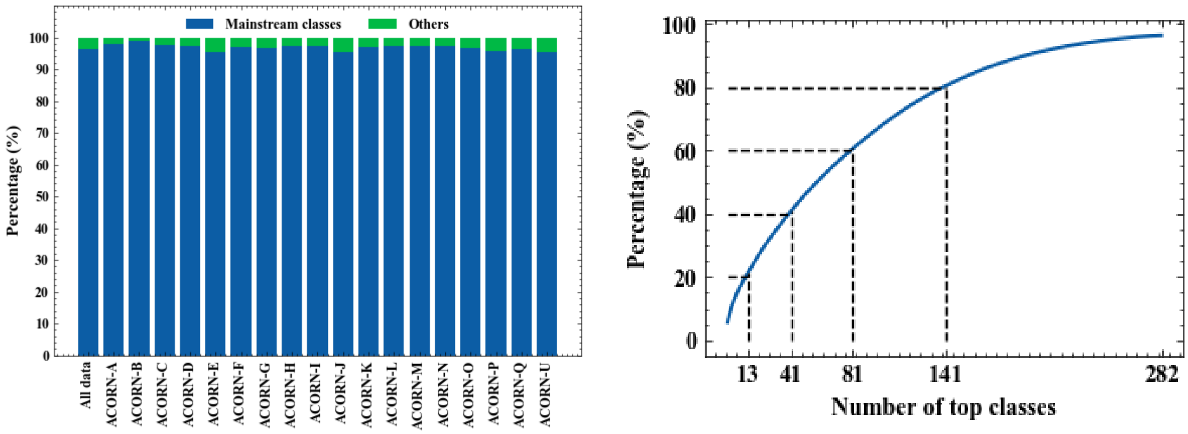
5.1 Load dictionary of DECP

The proposed three-step clustering approach produced a total of 458 final clusters. By allocating each sample to its corresponding final clusters according to the inheritance relationship between the three levels, we got the size of each final cluster and sorted them in descending order. Table 10 illustrates the number of top clusters corresponding to different percentages of samples and their minimum cluster size. Since we want to construct a load dictionary that represents mainstream DECPs, the cluster center of the top 282 clusters which account for 99.9 % of the samples were chosen. And for the other 176 clusters, their cluster centers were defined as low-frequency patterns and outliers, and labeled in the dictionary as "-1".

Table 10 - Size information of top clusters corresponds to different percentages of samples.

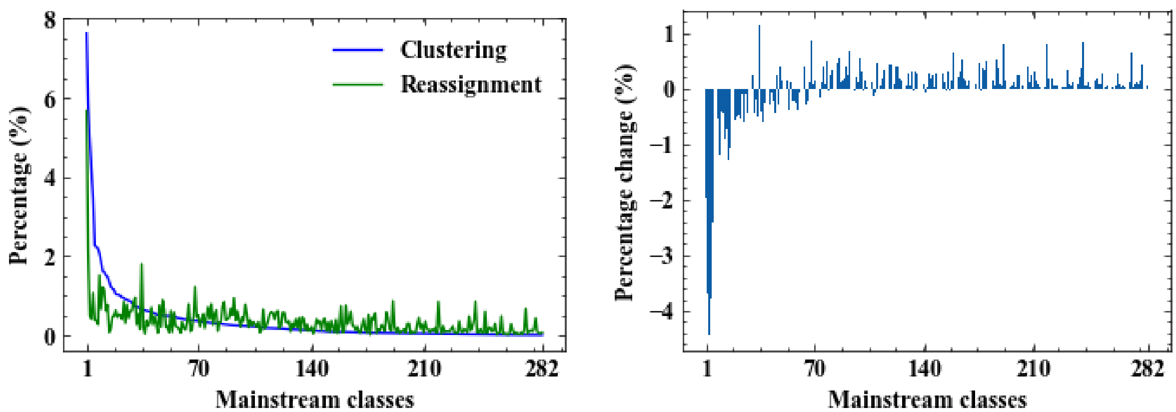
Percentage (%)	Number of top clusters	Minimum size of cluster
90.0	114	6220
95.0	152	3008
99.0	226	1094
99.9	282	204

Then, a reclassification of all samples was done by using the KNN model which took the 458 final cluster centers as training set. Each sample was assigned into one of the 283 classes which include the 282 mainstream classes and the 'dismissed class' ("-1"). According to the reclassification results (Figure 11), the new proportion of samples falling into the 282 mainstream classes is 96.46 %. This lower overall proportion, combined with the changes in the proportion of samples corresponding to each class shown in Figure 12, indicates that a substantial number of samples from the mainstream clusters are being reassigned to the 'dismissed class' of low-frequency pattern samples and outliers. The high proportion of samples corresponding to mainstream classes in both the initial clustering and reclassification results supports the validity of selected mainstream classes in the load dictionary. Figure A - 1 provides a sample visualization of the load dictionary. By means of performing manual checks, it's found that the load dictionary captures many minor differences between patterns (e.g., time periods and magnitude of value), while maintaining a relatively high level of distinction overall. In what remains of the analysis, the results of reclassification were used.



(a) (b)

Figure 11 - Sample percentage of mainstream classes (a) and top classes number (b) in reclassification results.



(a) (b)

Figure 12 - Sample percentage of mainstream classes (a) and their changes between clustering and reclassification (b).

After obtaining the load dictionary, the 282 mainstream classes were classified into different peak number categories (PNCs) according to their number of ECPks. The distribution of PNCs is shown in Figure 13, which indicates that the maximum number of peaks is 4 (Peak-4), and the distribution is close to that of the labeled dataset used for DLDA training. This implies that the trained semi-supervised clustering model effectively follows the sample pattern embodied in the training set for clustering.

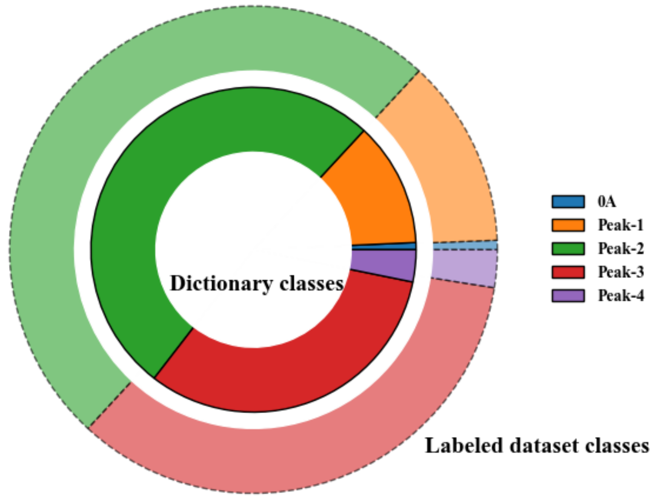
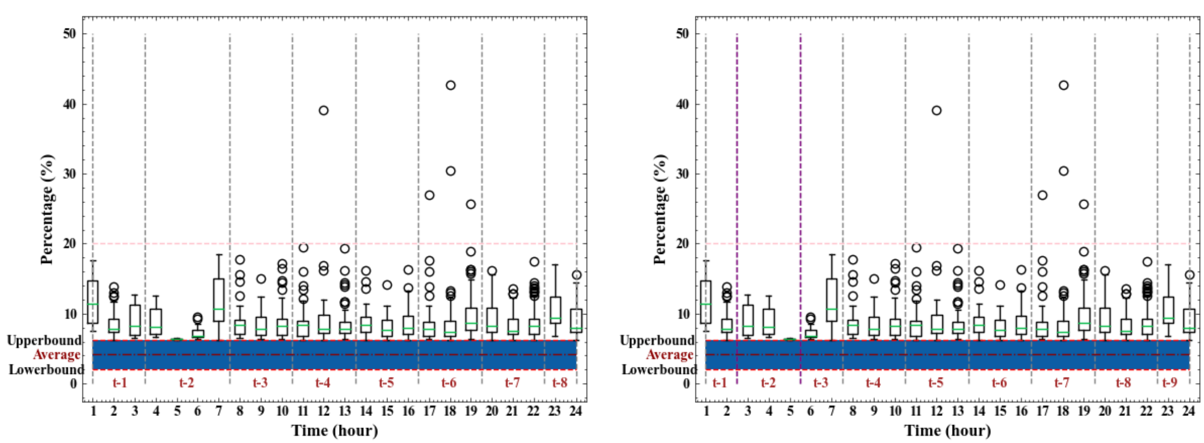


Figure 13 - Similarity of distributions of PNCs for mainstream load dictionary classes and labeled dataset classes.

The series of boxplots in Figure 14 show the distribution of peaks for the mainstream classes with respect to the hour; the blue band corresponds to the Average in Table 1. It's noted that in the vast majority of cases, the value of $p(t)$ is less than 20%. A more detailed inspection of the periods labeled as t-1 and t-2 in Figure 14(a) shows that the $p(t)$ varies considerably between 1 and 7 a.m.. What's more, the load sample statistics indicates that ECPks are much less likely to occur at 3-5 a.m. than at 1-2 a.m. and 6-7 a.m. (see Figure 17). This is important to note because it implies that there are substantial differences in electricity use habits. Thus, in order to highlight this, we re-partitioned this period into to 3, instead of 2 periods, as follows: 1-2 a.m. (t-1), 3-5 a.m. (t-2) and 6-7 a.m. (t-3). As a result, a new overall partitioning with a total of 9 time periods throughout the day was built (Figure 14(b)). The subsequent analysis was carried out using this new partitioning.



(a) (b)

Figure 14 - Time distribution of ECPks in mainstream load dictionary classes: (a) original partitioning and (b) re-partitioning of time.

5.2 Analysis of household ECB

Based on the reclassification results obtained earlier and the DEC data, a detailed analysis of London's household ECB is provided in this section.

5.2.1 Analysis of DEC

It's found that the $log(1 + DEC)$ conforms to a Gaussian distribution as shown in Figure 15, which is similar to that described by Kwac et al. [24] for household electricity consumption in California, USA; in both cases, a relatively concentrated household DEC distribution is observed.

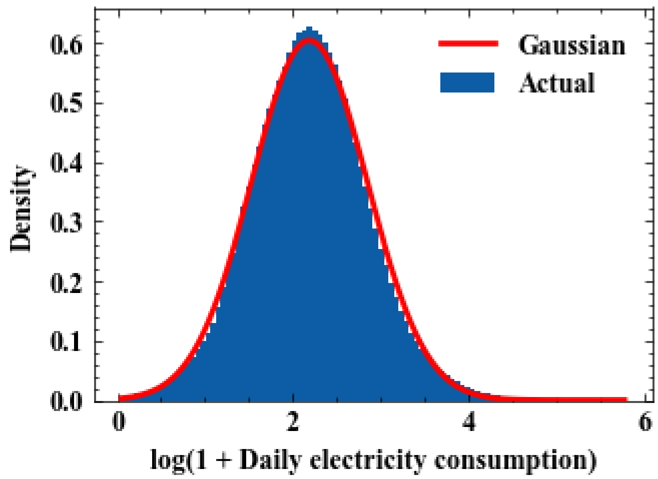


Figure 15 - Distribution of $log(1 + DEC)$ and corresponding Gaussian distribution fit.

Electricity consumption across the 18 Acorn groups is compared using the mean (M), standard deviation (Std), 25 % (P25), 50 % (P50) and 75 % (P75) percentile of DEC. As shown in Figure 16, from ACORN-A to ACORN-Q, DEC shows a roughly decreasing trend, which is consistent with the decreasing economic income trend, according to the relevant socio-demographic data. It is noteworthy that the Std of DEC for ACORN-A, D, E and J, which correspond to "Lavish Lifestyles", "City Sophisticates", "Career Climbers" and "Starting Out" respectively, is large, indicating some divergence in electricity consumption within these groups. For ACORN-U, no further analysis is undertaken here due to the lack of socio-demographic information of this group.

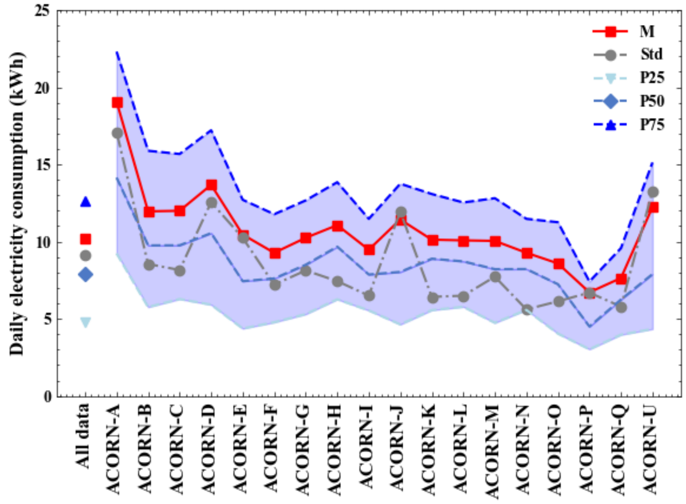


Figure 16 - Quantitative analysis of DEC for all data and Acorn groups.

5.2.2 Analysis of DECP

The frequency of ECPks during the 9 periods throughout the day was calculated based on the number of samples attributed to each mainstream class in the load dictionary to explore the general variation of electricity consumption within a day. In Figure 17, the vertical axis indicates the total frequency of samples with an ECPk during each of the 9 periods, which is understood as the probability of an ECPk occurring during that period. The results show that the dusk (t-7) and evening (t-8) periods have significant ECPk potential, with a probability around 0.5; this is perhaps not surprising as these are periods in which households are mainly involved in energy-intensive activities such as dining, entertaining and washing at home [23, 38-40]. Then it’s followed by daytime (t-4, t-5 and t-6) and late night (t-9), when the probability of a peak exceeds 0.2. Finally, in sleep time and early morning (t-1, t-2, t-3), the probability of ECPk is less than 0.1. The same trend is observed for virtually all Acorn groups. However, it is clear that the likelihood profile of ACORN-P, corresponding to the "Struggling Estates" group, differs significantly from the predominant pattern, with a considerable increase in the likelihood of a peak occurring during sleep time (>0.1) and a decrease in the probability of a peak at night.

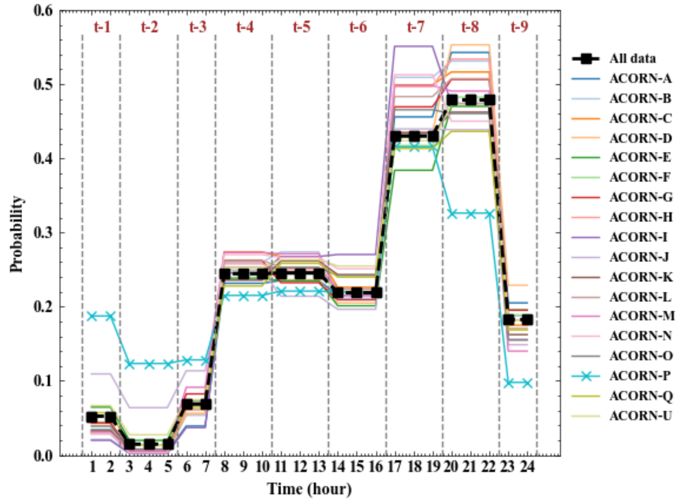
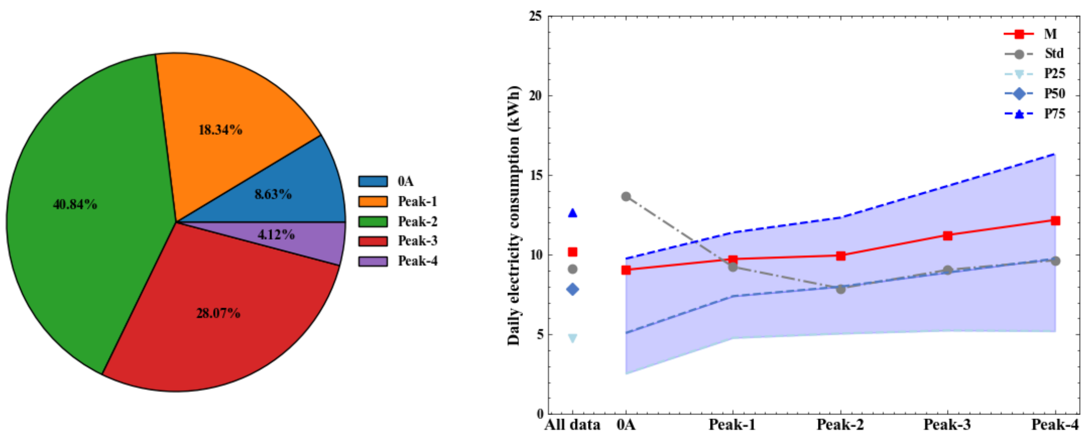


Figure 17 - Probability of ECPk occurrence in each time period for each Acorn group.

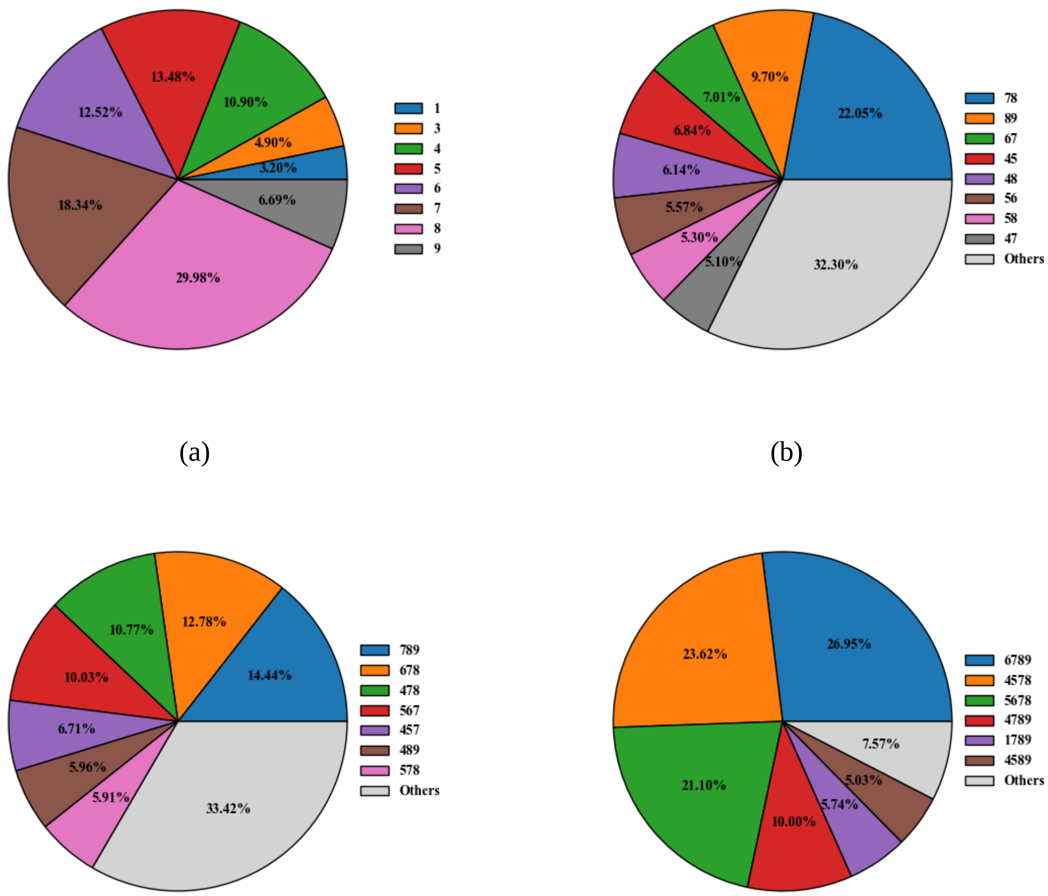
For the household DECP, due to the excessive number of mainstream classes, the high-level PNCs mentioned above were used for further analysis here. According to the number of samples attributed to each mainstream class, the proportion of samples and DEC of the PNCs was calculated as shown in Figure 18. Beyond our expectation, the percent of non-peak category 0A is 8.63%, higher than the smallest one Peak-4's 4.12%, and has the smallest M and the largest Std of DEC, which is not reflected in the time distribution of ECPk. This might suggest that a significant proportion of households have either high or low electricity consumption throughout the day. Furthermore, the trend of DEC increasing with the number of peaks is also observed in Figure 18 (b), though at a very low rate of increase. This appears to be in line with the common assumption that the longer the duration of the ECPks, the higher the electricity consumption becomes.



(a) (b)

Figure 18 - Sample percentage (a) and DEC (b) of PNCs.

Figure 19 and Figure 20 show the proportional breakdown of DECP samples according to the ECPk potential associated with each time period and the shape of their top 3 proportional groups of DECP samples, respectively, under the Peak-1 to Peak-4 categories. The results presented in these figures are an indication that the probability of ECPk for the different time periods is consistent with the results shown in Figure 17.



(a) (b)

(c) (d)

Figure 19 - Sample percentage of DECP represented by ECPk time periods under Peak-1 (a), Peak-2 (b), Peak-3 (c) and Peak-4 (d).

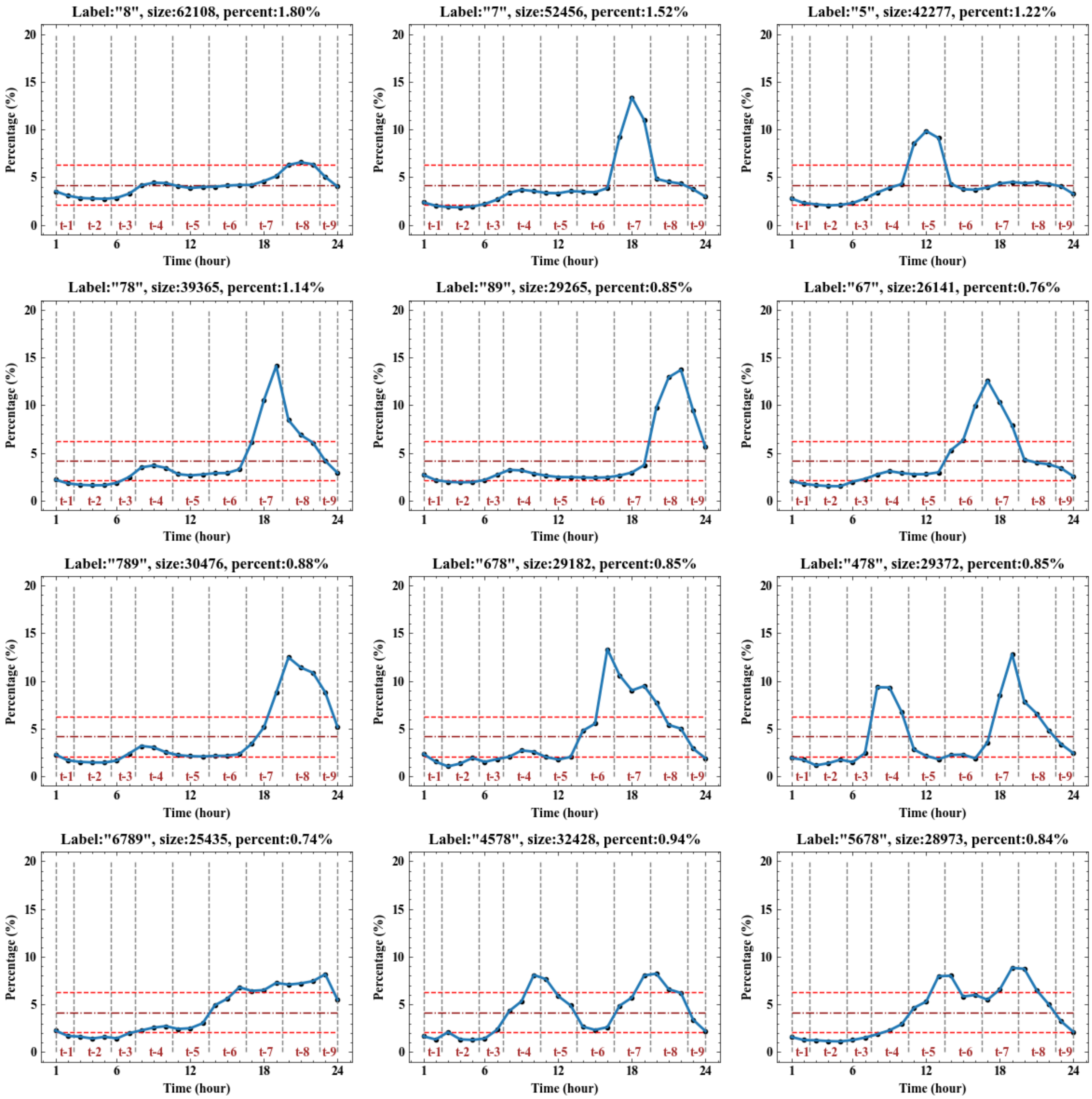


Figure 20 - Shape of top three sample percentage DECPs in each PNC. Figures from top to bottom represent Peak-1 to Peak-4 categories.

5.3 Similarity analysis of ECB among households

Based on the household DEC and DECP results obtained in the previous section, this section examines the similarity of ECB between households by clustering them. To ensure the representativeness of the results, households with more than 365 load samples were selected for analysis, which yields a total of 5310 households, which corresponds to 95.56% of the total number of households and still provides a socio-demographic composition consistent with that of the London population.

Firstly, to demonstrate the complexity of DECP variation within individual household datasets, the level of entropy (Equation (13)) is calculated for each of the load dictionary classes contained within the dataset of household load samples. The results are shown in Figure 21.

$$Entropy(x) = \sum_{c=1}^{C} -p_c \log(p_c) \qquad (13)$$

Where C is the number of load dictionary classes, and $p_c = \frac{N_c}{N}$ is the ratio between the sample size of class c (N_c) and the household (N).

Households in the dataset have an average of 124 DECPs ("-1" is considered a valid class). And the entropy values representing their variability are also more concentrated, which suggests that even the variability of DECP within individual household datasets is extremely complex. This phenomenon was also identified by Kwac et al. [24] and poses a significant challenge to the quality of the clustering of household DECP described below.

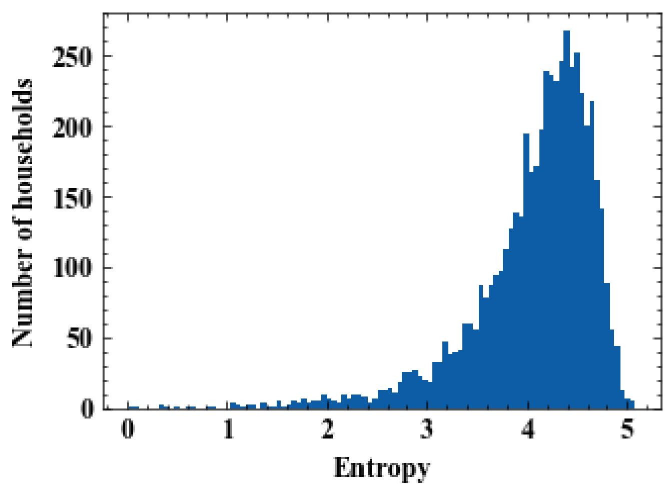
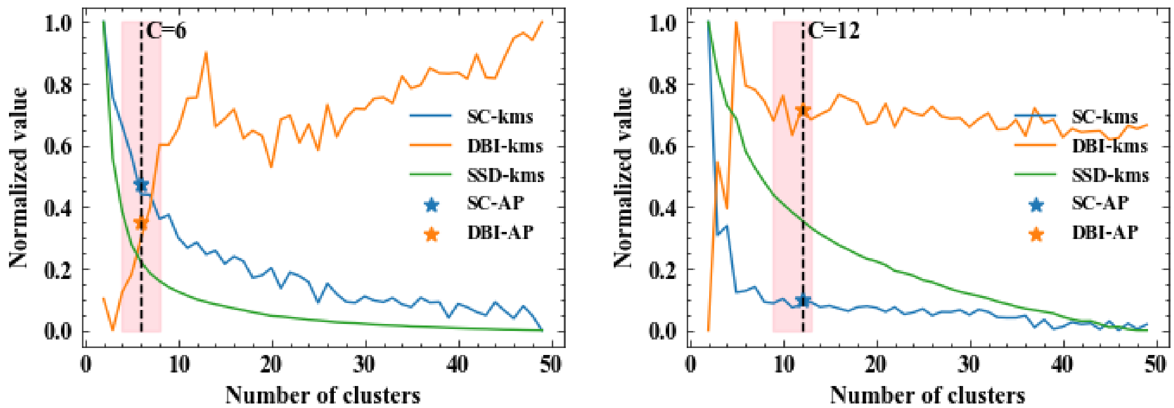


Figure 21 - Distribution of entropy for household DECPs.

5.3.1 Segmentation of households

In order to group households with similar ECB into the same category, clustering based on household DEC and DECP was carried out separately. The M, Std, P25, P50 and P75 of DEC were chosen as the features for household DEC clustering. The proportional breakdown of the 282 mainstream classes attributed to the load samples in each individual household dataset was chosen as the feature of household DECP clustering as well. The number of clusters was selected by comparing the results of SC, DBI and sum of squared distances of samples to their closest cluster center (SSD) yielded from K-means and AP clustering. Figure 22 shows the evaluation criteria of K-means clustering with the number of clusters set to 2 through to 49 and AP clustering results with the preference set to the minimum similarity input value to obtain the minimum number of clusters. For K-means, the appropriate number of clusters of DEC and DECP are between 4-8 and 9-13, respectively, according to the "elbow principle" of SSD in conjunction with the value of SC and DBI. The clustering results of AP all lie within these intervals. Considering the superior clustering performance of AP over K-means [21], the clustering results of AP were finally selected. That is, the number of clusters for household DEC and DECP was 6 and 12, respectively. The information of clustering results is listed in Table 11 and Table 12.



(a) (b)

Figure 22 - Optimal number of clusters based on clustering quality assessment for household DEC (a) and DECP (b).

Table 11 - SC and DBI of clustering results for household DEC and DECP using AP.

	DEC	DECP
SC	0.455	0.070
DBI	0.766	2.135
Number of clusters	6	12

Table 12 - Cluster size breakdown of clustering results for household DEC and DECP.

Cluster (DEC)	C1	C2	C3	C4	C5	C6
Size	110	389	21	1778	987	2025
Cluster (DECP)	C1	C2	C3	C4	C5	C6
Size	511	613	1206	72	80	495
Cluster (DECP)	C7	C8	C9	C10	C11	C12
Size	1613	143	298	17	38	224

The clustering results of household DEC show the 5310 households can be broadly classified into 6 classes; the corresponding class centers are shown in Figure 23. The four classes with lower DEC contain the majority of households and are also more concentrated in terms of DEC, which is consistent with the normally distributed DEC presented in Figure 15. Further analysis of class DEC-C3 was carried out, since it has the highest DEC; the results are shown in Figure 24. The top four Acorn groups for DEC-C3 correspond to ACORN-A, D, E, and J with the larger Std described above, while the boxplot of the percentage of PNCs for DEC-C3 households shows a significantly higher percentage of 0A and Peak-1 compared to the overall situation. Taking into account the fact that these households have a relatively favourable income profile, this could be attributed to the tendency of households in this category to have a high number of high-powered electrical appliances and often maintain high levels of utilisation throughout the day.

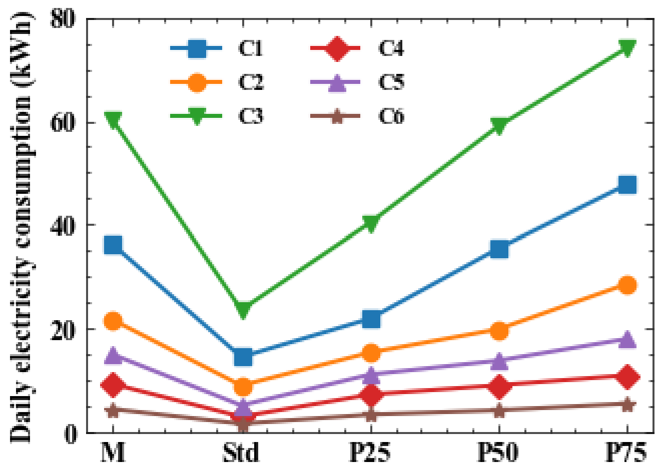
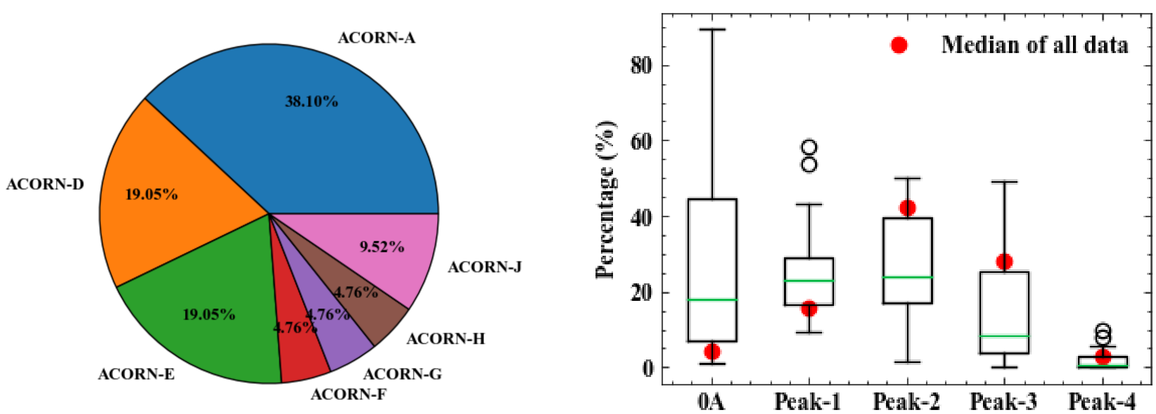


Figure 23 - Cluster centers of household DEC clustering results.



(a) (b)

Figure 24 - Distribution of Acorn groups (a) and percentage of PNCs (b) in class DEC-C3.

Figure 25 gives the distribution of household DECP clustering results in terms of the percentage of load samples allocated to the load dictionary classes in each cluster and the DECP entropy values of households in that cluster. As Figure 25 shows, the clustering reinforces the aggregation of households with similar DECP entropy values, but the high complexity of DECP variation within households allows the sample shares of DECP for similar households to still show some variation. Nonetheless, the clustering results reveal households in the same cluster show high similarity in certain DECPs.

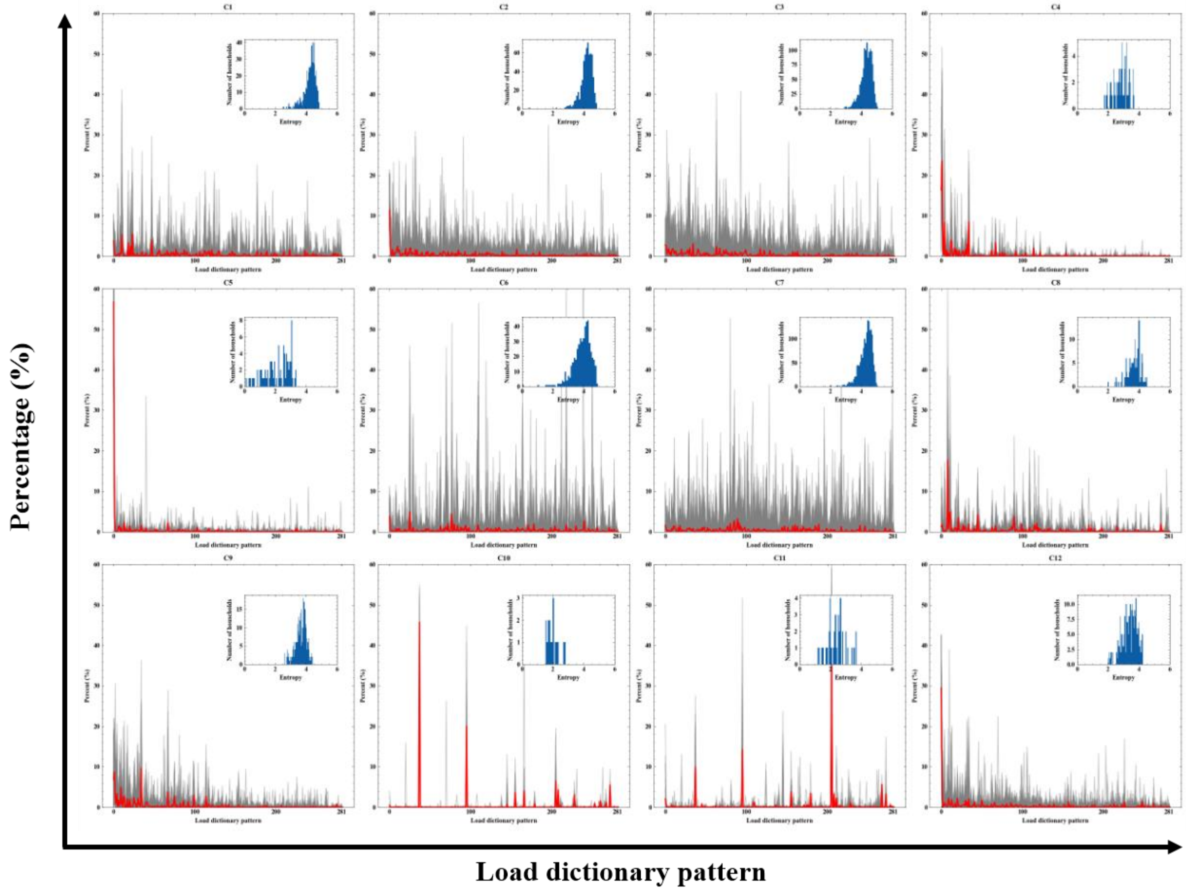


Figure 25 - Sample percentage of DECP; each panel corresponds to a cluster. For each individual cluster plot, the red curve represents the mean value, and the upper right corner subplot in blue shows its entropy distribution.

Based on the clustering results of household DEC and DECP, which were compared with the household Acorn groups, the mean of ARI and AMI are listed in Table 13. The results indicate there is no association between any of the three groupings. Therefore, a highly similar ECB household classification may be a combination of DEC categories representing similar capabilities and DECP categories representing similar operations.

Table 13 - Mean of ARI & AMI between the three results of household segmentation using the Acorn grouping, household DEC clustering and household DECP clustering.

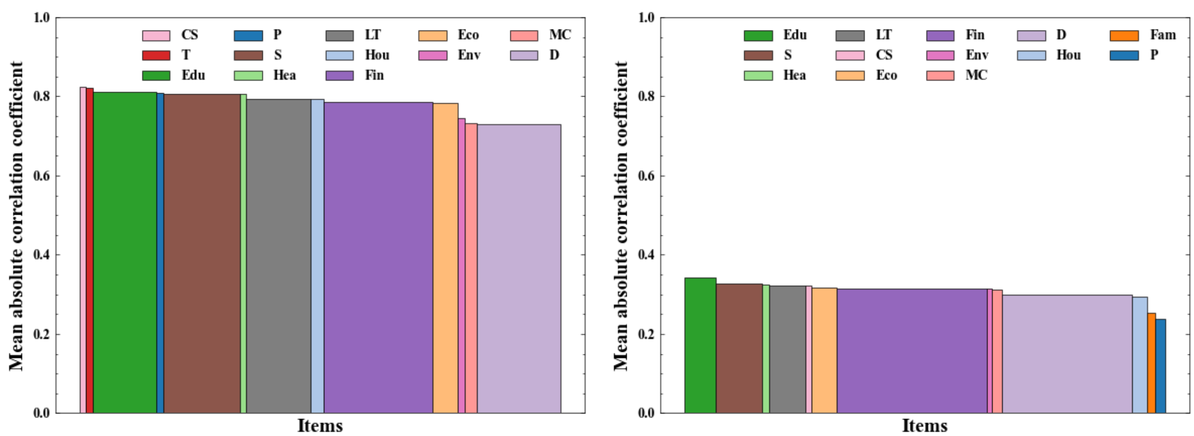
Groupings	DEC and DECP	DEC and Acorn	DECP and Acorn
Mean of ARI & AMI	0.008	0.018	0.007

5.3.2 Correlation between household properties and ECB

Although household Acorn groups do not match the exact segmentation of similar ECB households, it is feasible to use them as a single dataset with a large amount of electricity consumption data and then study the correlation between household properties and ECB. Therefore, this section explores the correlation between household ECB, as represented by the households' macro-DEC and DECP, and their socio-demographic characteristics.

The Acorn household classification is based on 15 major items and 84 subitems for a total of 826 options (as shown in Table A - 1). Correlations between each Acorn group's 826 option properties and its DEC and DECP were calculated respectively using the same features used for household clustering in the previous section. Based on the analysis, we found 445 and 822 household properties that were correlated with DEC and DECP features, respectively; 75 and 186 (with an overlap of 55) properties were found to have a strong correlation, as indicated by an absolute value of correlation coefficient (ACC) greater than 0.8. The results for the household properties strongly correlated with ECB are shown in Table A - 3, Figure A - 2 and Figure A - 3 in the Appendix.

Figure 26 shows the correlation between DEC and DECP, and major Acorn classification items. This is illustrated by using the mean ACC value for correlation between ECB and corresponding Acorn classification option properties. Due to the large number of patterns representing DECP and the fact that only certain patterns appear to be strongly correlated with particular types of household, they result in significantly lower values of DECP compared to the values obtained for DEC. It was found that there are no option properties which are strongly correlated with DEC for Family and Contact, and no option properties which are strongly correlated with DECP for Transport and Contact. Taken together, only Contact is observed not having a significant influence on the ECB of households, according to the analysis of linear relationship based on Pearson correlation coefficient. Besides, the correlation between the strongly ECB correlated option properties (Figure A - 4) reveals there are more independent factors influencing DECP than DEC, which accounts for the complex variability of DECPs in households.



(a) The 75 options strongly correlated with DEC; (b) The 186 options strongly correlated with DECP.

Figure 26 - Correlation between strongly ECB correlated household option properties and DEC (a) and DECP (b), represented by Mean ACC of major items.

5.4 Further potential applications

The analysis of household ECB helps in the fine-grained segmentation of household electricity customers and hence the formulation of different tariff strategies [19]. The paper presents a detailed analysis of the commonalities and characteristics of ECB across households, as well as its variation based on household DEC and DECP, and proposes a combination of DEC segmentation and DECP segmentation for electricity customers.

The DECP segmentation method based on a set of load dictionary classes may prove particularly useful as it effectively exploits the intra-day and inter-day variation characteristics of household ECB and captures underlying household electricity consumption habits. However, the strong emphasis on accuracy in this method comes at the cost of an increase in complexity. Higher level abstractions based on load dictionary classes, such as PNCs and pattern merging, can reduce the associated complexity while still providing a satisfactory description of household DECP at the macro level. Thus, it is important to keep in mind that choosing the most adequate ECB description method for different application scenarios can achieve the intended goal, even if this comes at the expense of some accuracy.

As we discussed in Section 3.1.1, the original data was resampled from half-hourly to hourly resolution. The hourly resolution was chosen due to considerations of significance of the differences between patterns, as well as difficulties associated with model training, but also the convenience of displaying the physical meaning of load samples and clustering results. However, moving towards higher resolutions remains an interesting avenue for further developments and, indeed, may prove necessary as more and more highly resolved data becomes available.

Another potential practical application for household ECB segmentation is household load forecasting. Based on the results of the analysis of the variability of individual household DECPs, it is clear that there is a wide variety of ECBs linked to individual households. Therefore, achieving high levels of accuracy when it comes to load forecasting for individual households might prove difficult if the full gamut of ECBs is not taken into account [41].

6. Conclusion

This study set out to address the current challenges posed by the ill-suited similarity metrics typically used for clustering of electrical load samples, and proposed a novel semi-supervised automatic clustering method which combines the supervised learning DLDA algorithm and the unsupervised learning AP clustering algorithm to overcome such problem. As a self-adapting metric learning method, DLDA+AP performs well on the LCL household electricity consumption dataset, obtaining much higher scores of ARI and AMI than traditional similarity metrics used in previous studies (73 to 78 percentage points higher; see Table 5 for comparison). It should be highlighted that the proposed method has a strong migration capability and the quality of the results is not tied to the particular choice of dataset. That is, it is possible to implement this method on a completely new dataset and obtain a high-quality automatic clustering solution adapted to the scenario due to the built-in similarity metric learning capabilities.

Based on the trained semi-supervised clustering model, the three-step clustering strategy and reclassification produced a load dictionary representing the typical DECP of households in the London area, which supported the analysis of household ECB from both DEC and DECP perspectives. In terms of ECB similarity between households, the clustering of DEC and DECP provided 6 and 12 categories respectively, and combining the two allowed for accurate targeting of a household ECB status. Further analysis involving the household properties provided by the Acorn household classification found that 206 option properties were strongly correlated with ECB, which included 14 of the 15 major items used for such classification.

As the deployment of smart grids technologies progresses, having the ability to quickly and effectively identify the potential grid requirements will be critical to their operation. In this regard, a detailed segmentation of the residential consumer base offers the possibility to create effective operation plans and prevent unforeseen peaks that threaten the stability of the grids. The residential sector already accounts for one of the largest shares of demand during peak times, and this is only likely to worsen as the electrification of other energy-intensive end uses gets under way. This only further highlights the importance of developing more adequate methods for the identification of customer energy consumption patterns so that the implementation of demand-side management strategies can fully leverage its potential and ensure the smooth operation of the future smart systems that will allow us to keep the lights on.

Acknowledgements

The authors gratefully acknowledge support from the Basic Science Center Program for Ordered Energy Conversion of the National Natural Science Foundation of China [Grant No. 51888103] and UK Research and Innovation through the Centre for Research into Energy Demand Solutions [Grant No. EP/R035288/1].

Appendix

Table A - 1 Items and subitems used in the Acorn household classification methodology [30].

Items	Abbr. of items	Number of subitems	Subitems	Number of options
Population	P	5	Age	8
			Geography	4
			Ethnicity	5
			Country of Birth	11
			Religion	7
Housing	Hou	5	House Type	5
			House Tenure	6
			House Size	5
			House Value	7
			Moving House	1
Family	Fam	3	Structure	7
			Children in household	4
			Household Size	4
Economy	Eco	4	Economic Activity	8
			NS Socio-Economic Classifications	9
			Social Grade	6
			Occupation	11
Education	Edu	6	Highest Level of Qualifications (Adults)	6
			England: Pupils at the end of KS1	4
			England: Pupils at the end of KS2	4
			England: Pupils at the end of KS4	2
			Scotland: Pupils in the S4 cohort	1
			Scotland: Pupils in the S5 cohort	2
Health	Hea	1	Behaviours & Lifestyle	7
Transport	T	4	Travel To Work	8
			Public Transport Accessibility Level	8
			Car Ownership	4
			Main Car Class	5
Marketing Channels	MC	2	Channels Received	12
			Future Responses	15
Finance	Fin	14	Household Annual Income	7
			Financial Attitudes	5
			Financial Situation	6
			Benefits	3
			Credit Cards	6
			Savings and Investments	14
			Loans	4

			Insurance and Pensions	7
			Financial Channel: Arrange Current Account	5
			Financial Channel: Arrange commoditised financial product	5
			Financial Channel: Arrange considered financial product	6
			Financial Channel: Manage Current Account	6
			Financial Channel: Manage Savings Account	6
			Expenditure per person per week	14
Digital	D	17	Internet Access: Frequency	4
			Internet Access: Usage in Last Week	5
			Digital Attitudes	6
			Technology at Home	3
			TV on Demand	3
			Mobile phone	1
			Smartphone Brand	8
			Tablet Devices	1
			Social Media Activity (at least weekly)	16
			Social Media Brands (used at least weekly)	18
			Number of apps on mobile phone (free or paid)	6
			Types of internet usage : Laptop or PC	45
			Types of internet usage : Mobile Phone	46
			Types of internet usage : Tablet / iPad	45
			Regularly research on the internet	40
			Purchased on the internet	40
			Sites regularly visited	85
Shopping	S	7	Preferred Supermarket	7
			Food Shopping	4
			Clothing & Footwear Stores	3
			Furniture & Fittings Stores	3
			Electrical Stores	3
			High Street Retailers	12
			Attitudes	12
Contact	C	1	Preferred Channel	4
Environment	Env	3	Environmental Groups	1
			Action	10
			Attitude	1
Community Safety	CS	1	Crime Survey for England	22
Leisure Time	LT	11	Daily Newspapers	10
			Magazines Read	14
			Charities	13
			Books Read	6
			Interests & Hobbies	22

			Visit Pubs for a Drink - Day	3
			Visit Pubs for a Drink - Evening	3
			Visit Pubs for a Meal - Day	3
			Visit Pubs for a Meal - Evening	3
			Restaurants - Most Often	3
			Holiday Destination/Type	12

Table A - 2 Evaluation criteria for clustering and classification.

Criteria	Equation	Value meaning better results
ARI[42]	$ARI(U,V) = \frac{RI-E[RI]}{max(RI)-E[RI]}$ (test the similarity between U and V) where: $\cdot U,V$ denote label assignments, $\cdot RI(U,V) = \frac{a+b}{C_2^N}$ is the rand index between U and V, $\cdot a$ is the number of pairs of elements in the same class both in U and V, $\cdot b$ is the number of pairs of elements in different classes both in U and V, $\cdot C_2^N$ is the total number of pairs of elements (without ordering) $\cdot E[RI]$ is the expectation of RI.	Value closer to 1 (ARI $\in [-1,1]$)
AMI[43]	$AMI(U,V) = \frac{MI-E[MI]}{max(H(U),H(V))-E[MI]}$ (test the agreement between U and V) where: $\cdot U,V$ denote label assignments, $\cdot MI(U,V) = \sum_{i=1}^{\lvert U\rvert}\sum_{j=1}^{\lvert V\rvert} p\,(i,j) log\,(\frac{p(i,j)}{p(i)p'(j)})$ is the mutual information between U and V, $\cdot H(U) = \sum_{i=1}^{\lvert U\rvert} p\,(i) log\,(p(i))$ is the entropy of U, $\cdot E[MI]$ is the expectation of MI.	Value closer to 1 (AMI $\in [-1,1]$)
SC[44]	$SC = \frac{1}{N}\sum_{i=1}^{N}\frac{b_i-a_i}{max(a_i,b_i)}$ (from the perspective of sample) where: $\cdot N$ is the number of samples, $\cdot a_i$ is the average distance from sample i to other samples in the same class, $\cdot b_i$ is the average distance from sample i to other samples in the nearest different class	Value closer to 1 (SC $\in [-1,1]$)
DBI[45]	$DBI = \frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j} R_{ij}$ (from the perspective of cluster) where: $\cdot K$ is the number of clusters, $\cdot R_{ij} = \frac{s_i+s_j}{d_{ij}}$ is the similarity between cluster C_i and cluster C_j,	Value closer to 0 (DBI $\in [0,+\infty)$)

	$\cdot s_i$ is the average distance between each sample and cluster centriod in C_i, $\cdot d_{ij}$ is the distance between the cluster centriod of C_i and C_j	

Table A - 3 Acorn classification options strongly correlated (ACC>0.8) with either household DEC or DECP features.

Items	Subitems	Options	DEC (M, Std, P25, P50, P75)	DECP (load dictionary class)
Population	Age	Age 18-24		181
		Age 35-49		-274
	Geography	Scotland		216
	Ethnicity	Black		47,148
	Religion	Jewish	M,Std,P75	
Housing	House Type	Terraced house		42
	House Tenure	Mortgaged		-118
	House Size	Number of Beds : 2		6
		Number of Beds : 4		139,-147
		Number of Beds : 5 plus	M,P75	
	House Value	House Value up to 100k		118
		House Value 500k-750k		233
		House Value 1m+	M,Std,P75	
Family	Structure	Couple family no children		-118
	Household Size	Household size : 1 person		269
		Household size : 3-4 persons		-269
Economy	Economic Activity	Employee Full-Time		16,-274
	NS Socio-Economic Classifications	Higher managerial, administrative and professional occupations	M,Std,P75	-147,224
		Lower managerial, administrative and professional occupations		16,94,-147,224
		Routine occupations		147
	Social Grade	A	M,Std,P25,P50,P75	232
		B	P75	94,-147,224
		D	-P75	-94,118,147,-224
	Occupation	Director / Managerial		146
		Professional		16
		Retired		101
Education	Highest Level of Qualifications (Adults)	No formal qualifications		147
		Degree or higher degree	Std,P75	224
	England: Pupils at the end of KS1	Achieving expected level in writing	M,P75	224
	England: Pupils at the	Achieving expected level in		224

	end of KS2	reading		
		Achieving expected level in writing	M,P75	224
		Achieving expected level in maths	M,P75	224
		Achieving expected level in reading, writing and maths	M,Std,P50,P75	224,245
	England: Pupils at the end of KS4	Achieving 5+ A*-C at GCSE or equivalent	M,P50,P75	-6,-147,224,232
		Achieving 5+ A*-C (including English and Maths) at GCSE or equivalent	M,P50,P75	-147,224
	Scotland: Pupils in the S4 cohort	Attained 5 awards at SCQF level 5 and above	M,P50,P75	94,-147,224,232,238,245
	Scotland: Pupils in the S5 cohort	Attained 3 awards at SCQF level 6 and above	M,P50,P75	94,-147,224,232,238,245
		Attained 5 awards at SCQF level 6 and above	M,Std,P50,P75	-147,224,232,238,245
Health	Behaviours & Lifestyle	Takes regular exercise	M,Std,P75	-147,224
		Eats fruit 3 or less days per week		118,-224
		Eats vegetables 3 or less days per week		-224
Transport	Main Car Class	Luxury or Executive	M,Std,P75	
Marketing Channels	Channels Received	Email		-147
		Cinema Advertising	P75	-147
	Future Responses	Mail - Addressed to you by name	P75	94,224
		Newspaper / Magazine Adverts		224
Finance	Household Annual Income	￡0-￡20,000		118,147,-224,269
		￡60,000-￡80,000	P75	94,-118,-147,224,-269
		￡80,000-￡100,000	M,Std,P75	-147,224,-269
		￡100,000+	M,Std,P50,P75	-147,224
		Average Household Income	M,Std,P75	-147,224,-269
	Financial Attitudes	I am very good at managing money		50
		Financial security after retirement is your own responsibility		50

	Financial Situation	Saving		-147,224
		Not saving		147,-224
		Saving a lot	M,P75	-147,224
		Saving a little		-118
		Just managing to make ends meet		147,-224
	Benefits	Job Seeker's Allowance		118
		Disability Living Allowance		118
		Income Support		118
	Credit Cards	Has credit card		-118,-147,224
		Has 2+ credit cards		94,-147,224
		Spent £500+ in last month on a credit card	M,P75	-147,224
		Uses credit card 6+ times per month	M,P75	-147,224
		Usually makes minimum payment on card		137
		Always pays credit card balance in full		-147,224
	Savings and Investments	Has savings account		-118
		Has instant access account		-118
		Has Stocks and Shares ISA	M,P50,P75	-147,224,232
		Has Unit Trusts	M,Std,P75	
		Has stocks and shares	M,P50,P75	-147,224,232
		Has investment bonds	P75	-147,224
		Has Investments	M,P75	-147,224,232
		Value of investments £25,000+	M,Std,P25,P50,P75	232
		Savings value £1 - £500		168
		Savings value £10,000+	M,P75	-147,224,232
	Loans	Has 2+ loans		189
		Unsecured debt greater than £15,000		-118
	Insurance and Pensions	Has Private Health Care	M,Std,P75	224
		Has Company Health Care		16,146,224
		Has Life Assurance		22
		Has life protection policy		-181
		Has pension scheme organised through company		-274
		Has pension scheme organised personally		16,94,-118,146,-147,224,-269
		Plans to use other investments for retirement		-147,224

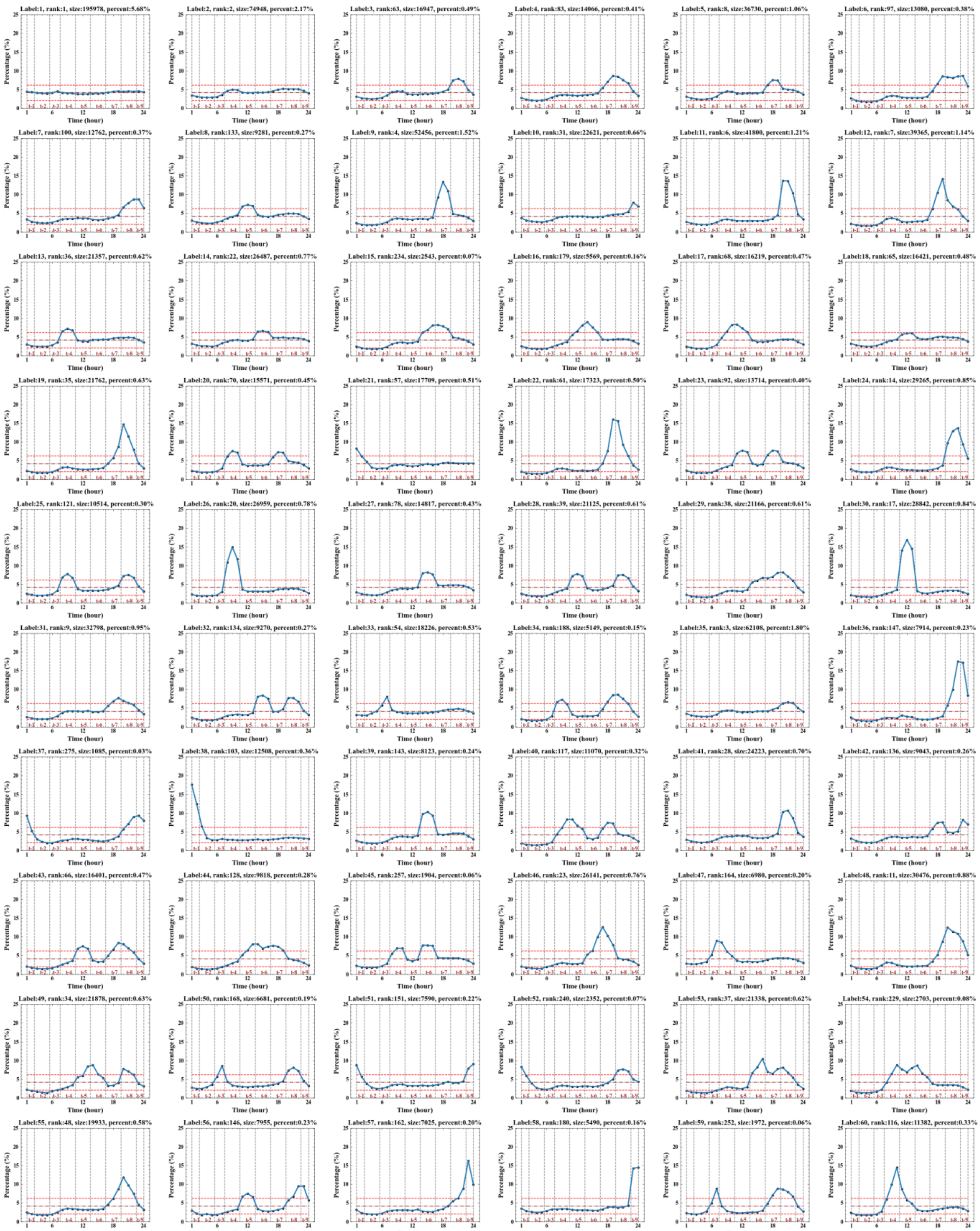
	Financial Channel: Arrange commoditised financial product	Online		-118
		By post		16,94,224,245
	Financial Channel: Arrange considered financial product	Online		94,-118,-147,224,-269
		By phone		16,-118,146,-147,224,-269
		By post		-118,-147,224,-269
		Used price comparison site		16
		Used IFA		16,94,-118,146,-147,224,-269
	Financial Channel: Manage Current Account	In branch		-16,-224,269
		Online		16
		Online at home or work		16,-147,224
	Expenditure per person per week	Total Expenditure	Std	-147,-192,224,-269
		Clothing and footwear		-147,224,-269
		Furnishings, household equipment and routine maintenance	M,P75	-6,-147,224
		Health		-118,-147,224
		Transport		-118,-147
		Recreation and Culture		-147
		Restaurants and hotels		-147,224,-269
		Miscellaneous goods and services		-147,224,-269
		Total Online Expenditure		-118,146,-147,224,-269
Digital	Internet Access: Usage in Last Week	8-19 hours		-147
	Technology at Home	Has a smartwatch, fitness band or payment band		224
	Smartphone Brand	Samsung		140
	Tablet Devices	Has Tablet (e.g. iPad, Samsung Galaxy Tab, Sony Xperia)		-147,224,-269
	Social Media Activity (at least weekly)	Update your status/tell people what you are up to/tell people what's happening		-227
	Types of internet usage : Laptop or PC	Check stocks/shares/investments		16,94,146,-147,224
		Gambling/Betting		-232
	Types of internet usage : Mobile Phone	Listen to the radio		189
	Types of internet usage :	Enter competitions		-269,-274

	Tablet / iPad	Use GPS services (e.g. Maps)		-192
		Send and receive personal emails		-147,224,-269
		Listen to the radio		-12,-192
		Take pictures/video clips	M,Std	
		Source info on products/services via Internet		-269
		Download music via the Internet		-97
		Locate places/shops/businesses via the Internet		-147
		Manage personal finances (banking) via the Internet		224,-269
		Post ratings and reviews		-222,268,-275
		Access news via news site (other than online newspapers)		-269
		Online auctions (e.g. ebay)		-269
		Read newspapers online		16,224
		Read magazines online	Std	
		Make an online purchase		16,-118,-269
	Regularly research on the internet	Savings/Investments		-147,224
		Household utilities (e.g. Gas, Electricity)		-118
		Furniture		58,189
		Sports equipment		-147
		Mortgages		189
		Other Insurance (e.g. Life, Travel)		16,94,-118,224,-269
		Books		224,-269
		Hotel reservations		16,94,-118,146,-147,224,-269
		Airline tickets	Std	
		Holidays		-118,-147,224
	Purchased on the internet	Credit Cards		-37
		Household utilities (e.g. Gas, Electricity)	Std	224
		Other Insurance (e.g. Life, Travel)		16,94,146,224
		Household electrical products	Std	
		Books		-147
		Hotel reservations		224
		Sports and Leisurewear/trainers		-222
		Holidays		16,146

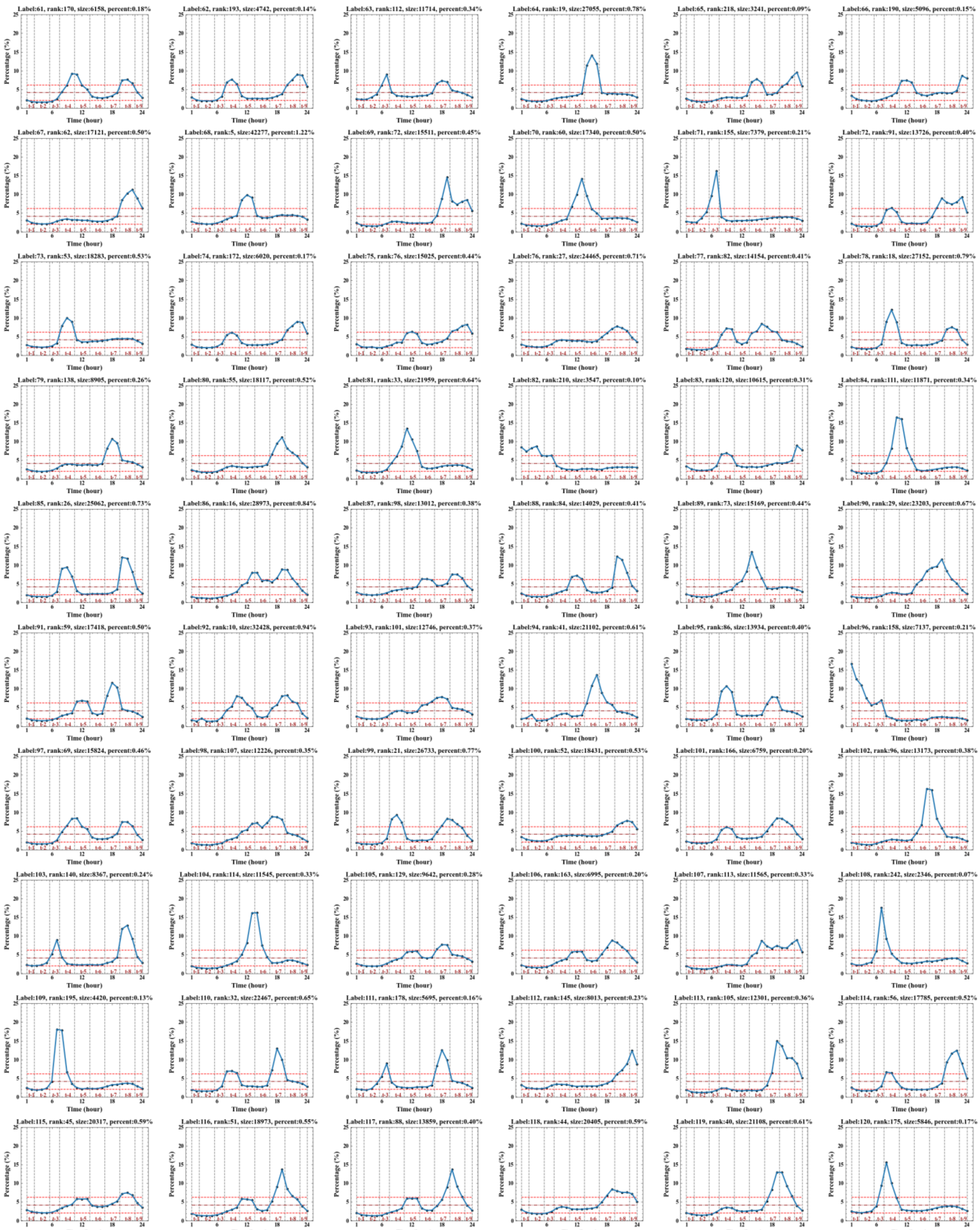
		None of these		-147,224
	Sites regularly visited	Argos	-M,-P50,-P75	-94,147,-224,-245
		bbc.co.uk		16
		Boots.com		189
		eHow.com		-259
		johnlewis.com	M,Std,P75	-147,224
		Lastminute.com	M	
		Marks and Spencer	P75	-118,-147,224
		Moneysavingexpert.com	M,Std,P75	-118,-147,224
		Moneysupermarket.com		-118
		Myvouchercodes		-179
		Net-a-Porter	M	
		Notonthehighstreet.com		16
		Play.com	-M	
		Quidco		16
		Ticketmaster	M,Std	
		Tripadvisor		16,94,146,224
Shopping	Preferred Supermarket	Asda	-M,-P75	-238,-245
		M & S	M,Std,P50,P75	94,-147,224,245
		Waitrose	M,Std,P75	224,245
	Food Shopping	Fairtrade	M,Std,P50,P75	232
		Premium Ranges	M,P75	-147,224
	Clothing & Footwear Stores	Premium	M,Std,P75	-147,224
		Value	-M,-Std,-P50,-P75	-224,-245
	Furniture & Fittings Stores	Premium	M,P75	-147,224
	Electrical Stores	Premium	M,Std,P75	-147,224
		Mass Market		-118
		Value	-M,-P25,-P50,-P75	-94,147,-224,-232
	High Street Retailers	McDonalds		-227
		Poundland		-16
	Attitudes	I am prepared to pay more for products that make life easier		224
		It's worth paying extra for quality goods	M,P75	-147,224
		I tend to go for premium rather than standard goods/services	M,Std,P75	224
		I only shop at supermarkets that sell good quality fresh food		50
		I check a number of sources before making a significant purchase		-147,224
Environme	Action	Rarely keep the tap running		-118

nt		while brushing teeth		
		Rarely leave the heating on when out for a few hours	-P25	-232,-259
Community Safety	Crime Survey for England	Breakdown of family	M,P50,P75	
		Being physically attacked by strangers		118
		Being raped		147
Leisure Time	Daily Newspapers	Daily Star		118,147,-224
		The Times	M,Std,P50,P75	
		Daily Telegraph	M,P75	139,-147,224
	Magazines Read	Travel		259
	Charities	Overseas Development		224
		Regularly donate to charity	M,P50,P75	-147,224,232
	Books Read	History/Biography	M,Std,P75	
	Interests & Hobbies	Antiques or Fine Art	M,Std,P75	
		Foreign Travel	M,P25,P50,P75	232,259
		Gambling		-16,-94,118,-146,-224
		Healthy Eating		52,63,119,187,193,199
		Organic Foods	M,Std	
		Reading Books		-262
		TV		242
	Restaurants - Most Often	Premium		16,146,224
	Holiday Destination/Type	USA / Canada	M	
		Caribbean		224
		Africa	M,Std	
		Activity / Outdoor Sports	M,P75	-6,139,-147,224
		Package		259

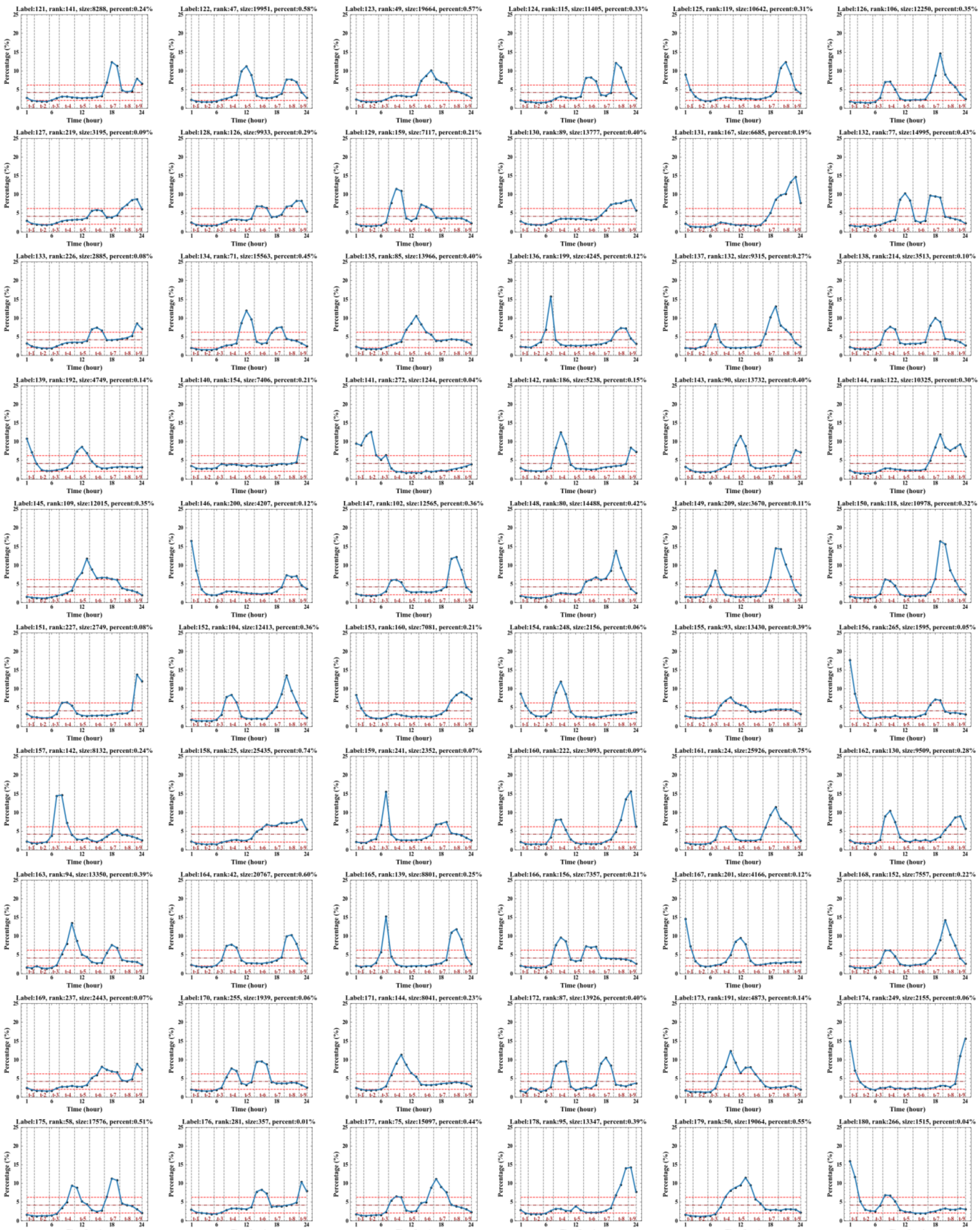
Notes: “-” means negative correlation. For example, “-M” means the option is negatively correlated with M (the mean) of DEC, or “-6” means that the option is negatively correlated with the load dictionary class 6 of DECP.


Label:1, rank:1, size:195978, percent:5.68%
Label:2, rank:2, size:74948, percent:2.17%
Label:3, rank:63, size:16947, percent:0.49%
Label:4, rank:83, size:14066, percent:0.41%
Label:5, rank:8, size:36730, percent:1.06%
Label:6, rank:97, size:13080, percent:0.38%
Label:7, rank:100, size:12762, percent:0.37%
Label:8, rank:133, size:9281, percent:0.27%
Label:9, rank:4, size:52456, percent:1.52%
Label:10, rank:31, size:22621, percent:0.66%
Label:11, rank:6, size:41800, percent:1.21%
Label:12, rank:7, size:39365, percent:1.14%
Label:13, rank:36, size:21357, percent:0.62%
Label:14, rank:22, size:26487, percent:0.77%
Label:15, rank:234, size:2543, percent:0.07%
Label:16, rank:179, size:5569, percent:0.16%
Label:17, rank:68, size:16219, percent:0.47%
Label:18, rank:65, size:16421, percent:0.48%
Label:19, rank:35, size:21762, percent:0.63%
Label:20, rank:70, size:15571, percent:0.45%
Label:21, rank:57, size:17709, percent:0.51%
Label:22, rank:61, size:17323, percent:0.50%
Label:23, rank:92, size:13714, percent:0.40%
Label:24, rank:14, size:29265, percent:0.85%
Label:25, rank:121, size:10514, percent:0.30%
Label:26, rank:20, size:26959, percent:0.78%
Label:27, rank:78, size:14817, percent:0.43%
Label:28, rank:39, size:21125, percent:0.61%
Label:29, rank:38, size:21166, percent:0.61%
Label:30, rank:17, size:28842, percent:0.84%
Label:31, rank:9, size:32798, percent:0.95%
Label:32, rank:134, size:9270, percent:0.27%
Label:33, rank:54, size:18226, percent:0.53%
Label:34, rank:188, size:5149, percent:0.15%
Label:35, rank:3, size:62108, percent:1.80%
Label:36, rank:147, size:7914, percent:0.23%
Label:37, rank:275, size:1085, percent:0.03%
Label:38, rank:103, size:12508, percent:0.36%
Label:39, rank:143, size:8123, percent:0.24%
Label:40, rank:117, size:11070, percent:0.32%
Label:41, rank:28, size:24223, percent:0.70%
Label:42, rank:136, size:9043, percent:0.26%
Label:43, rank:66, size:16401, percent:0.47%
Label:44, rank:128, size:9818, percent:0.28%
Label:45, rank:257, size:1904, percent:0.06%
Label:46, rank:23, size:26141, percent:0.76%
Label:47, rank:164, size:6980, percent:0.20%
Label:48, rank:11, size:30476, percent:0.88%
Label:49, rank:34, size:21878, percent:0.63%
Label:50, rank:168, size:6681, percent:0.19%
Label:51, rank:151, size:7590, percent:0.22%
Label:52, rank:240, size:2352, percent:0.07%
Label:53, rank:37, size:21338, percent:0.62%
Label:54, rank:229, size:2703, percent:0.08%
Label:55, rank:48, size:19933, percent:0.58%
Label:56, rank:146, size:7955, percent:0.23%
Label:57, rank:162, size:7025, percent:0.20%
Label:58, rank:180, size:5490, percent:0.16%
Label:59, rank:252, size:1972, percent:0.06%
Label:60, rank:116, size:11382, percent:0.33%
Percentage (%)
Time (hour)

(a) Label 1 – 60

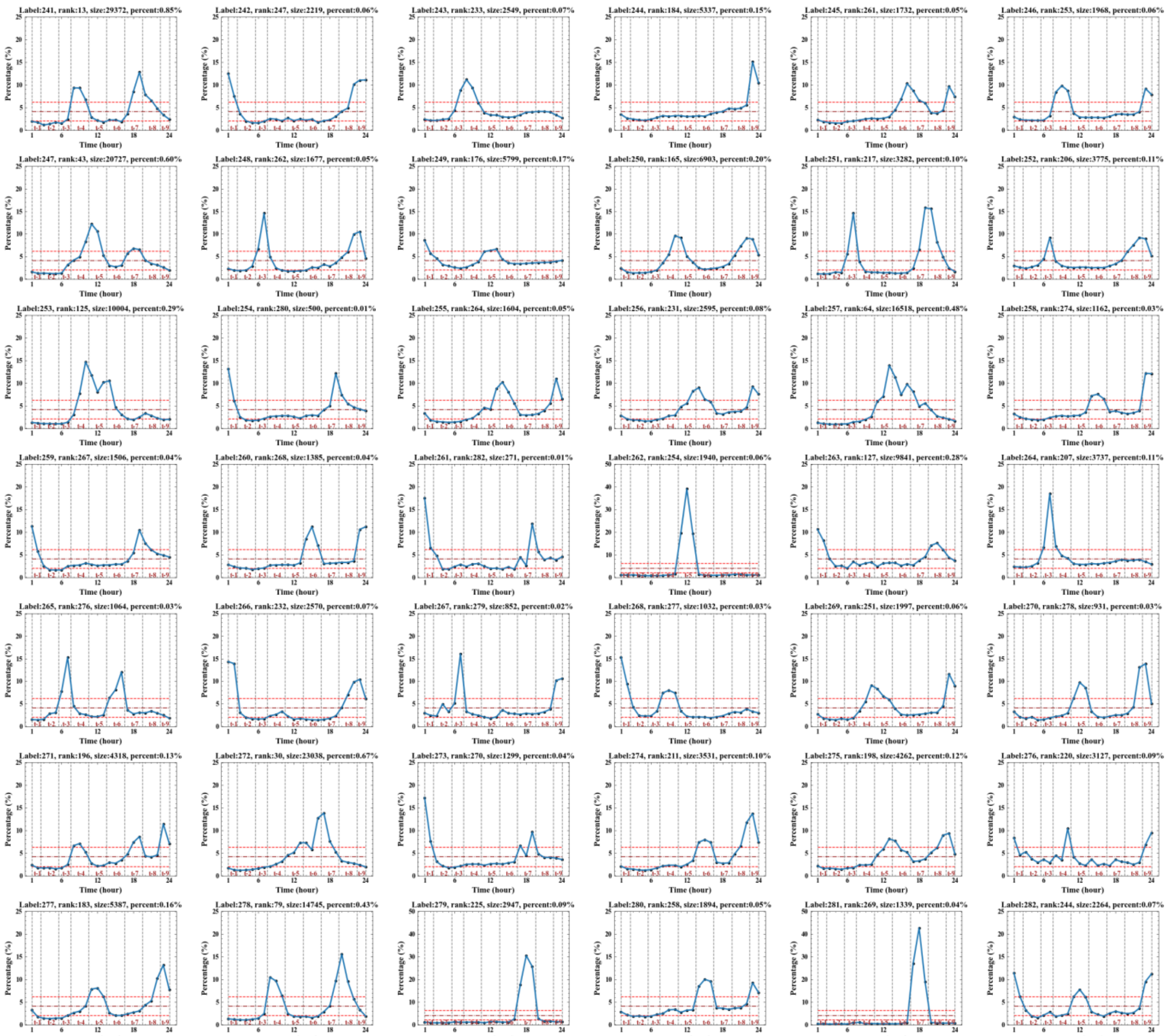


(b) Label 61-120



(c) Label 121-180

(d) Label 181-240



(e) Label 241-282

Figure A - 1 Visualization of load dictionary.

("Label" means the order of the class in load dictionary; "rank" means the order of sample size of the class in load dictionary; "size" means the sample size of the class; "percent" means the sample percentage of the class. All the data comes from reclassification results.)

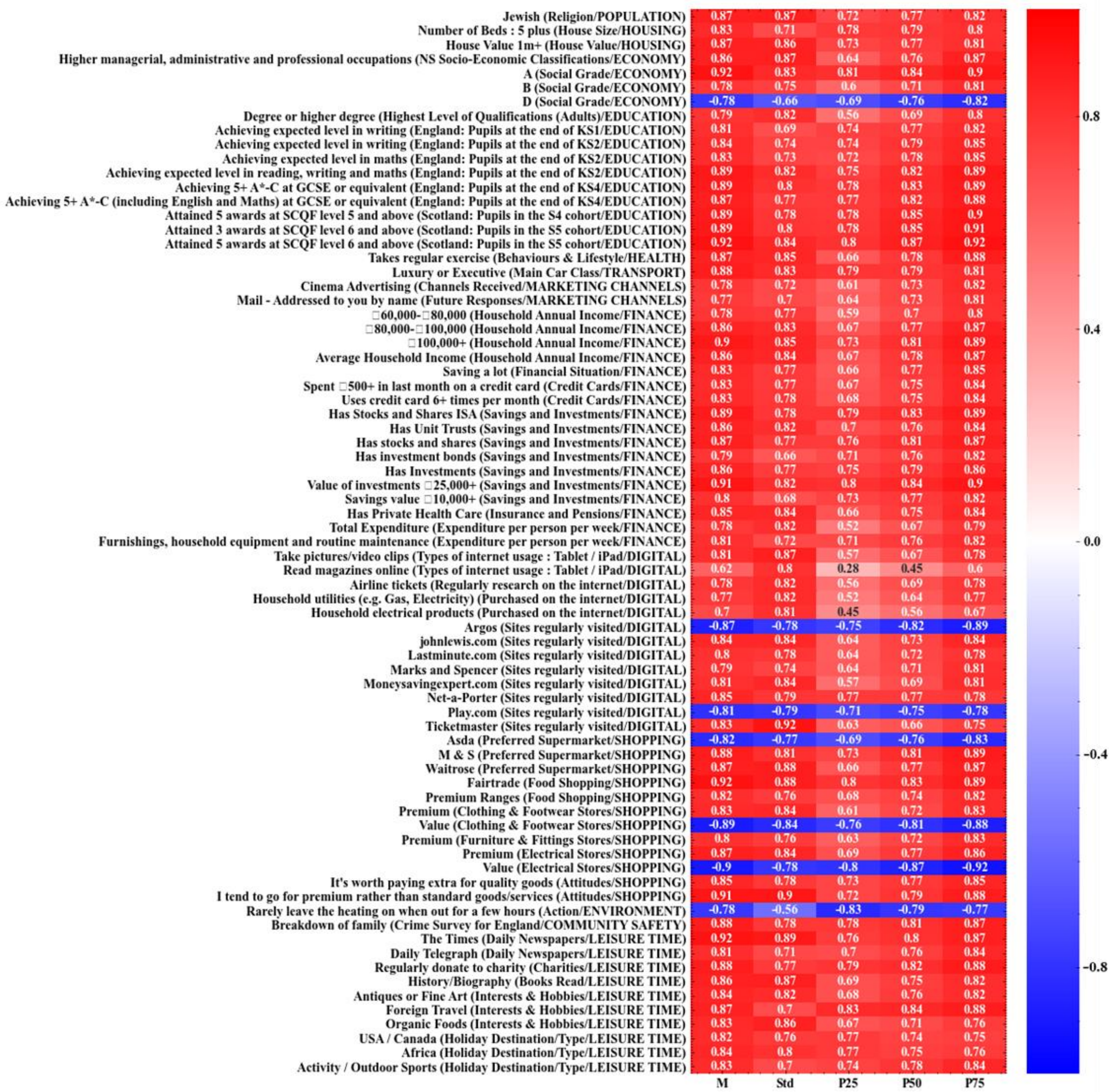


Figure A - 2 Heatmap of correlation coefficients between proxy features of DEC and Acorn option properties with ACC > 0.8.

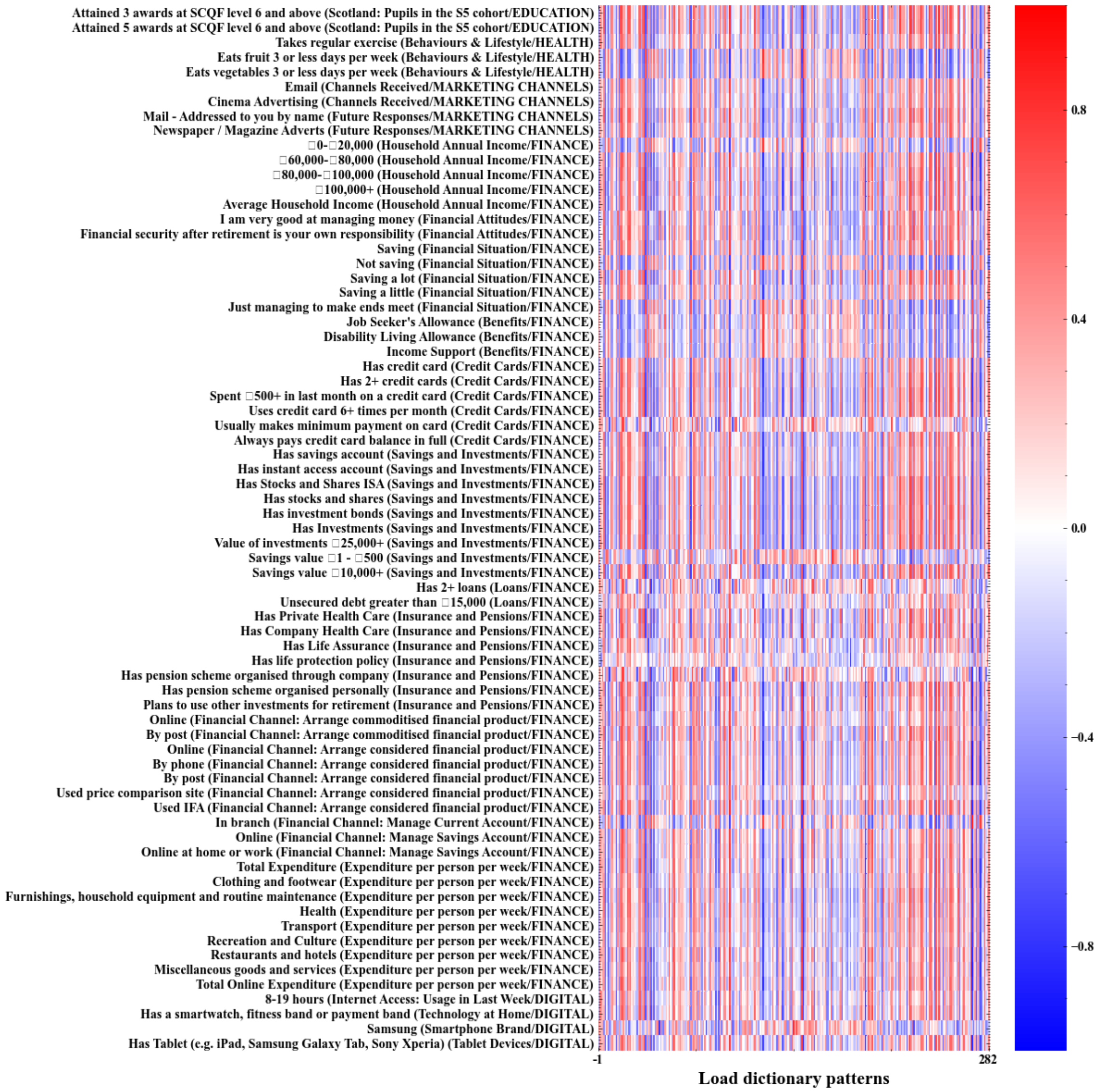
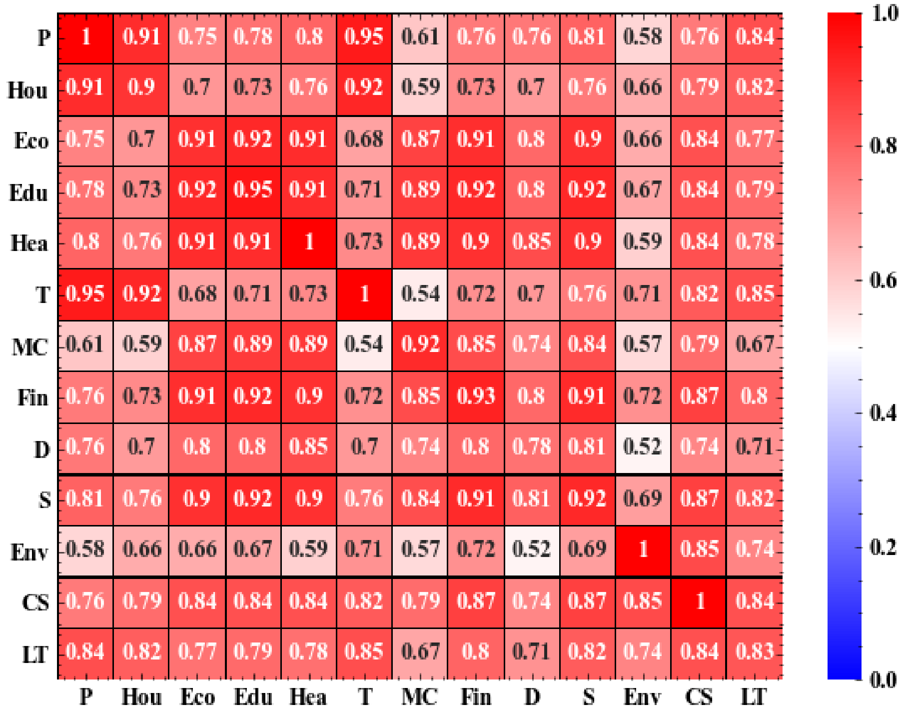
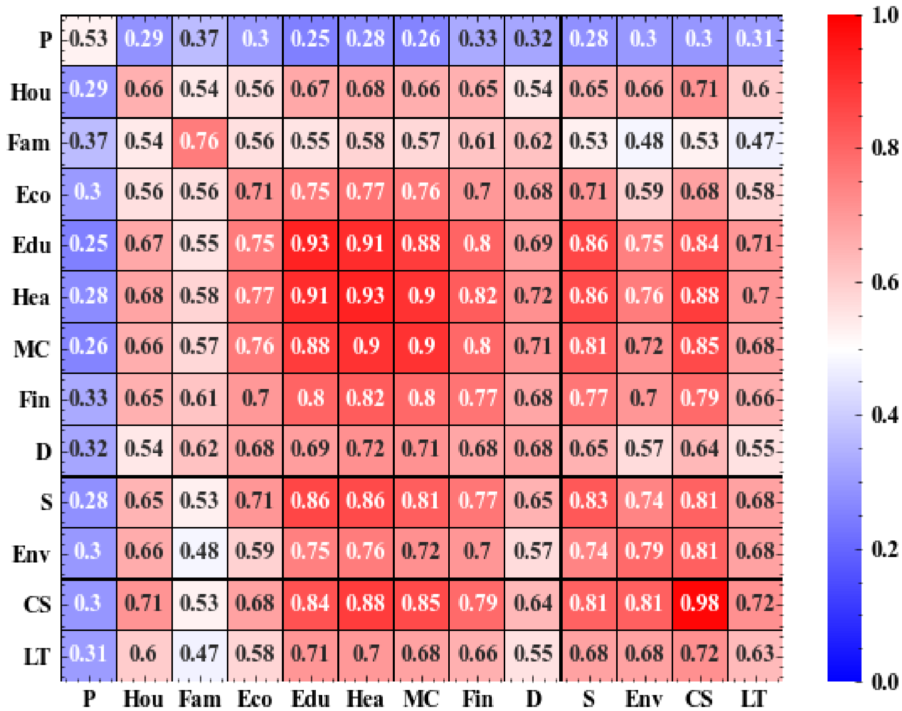


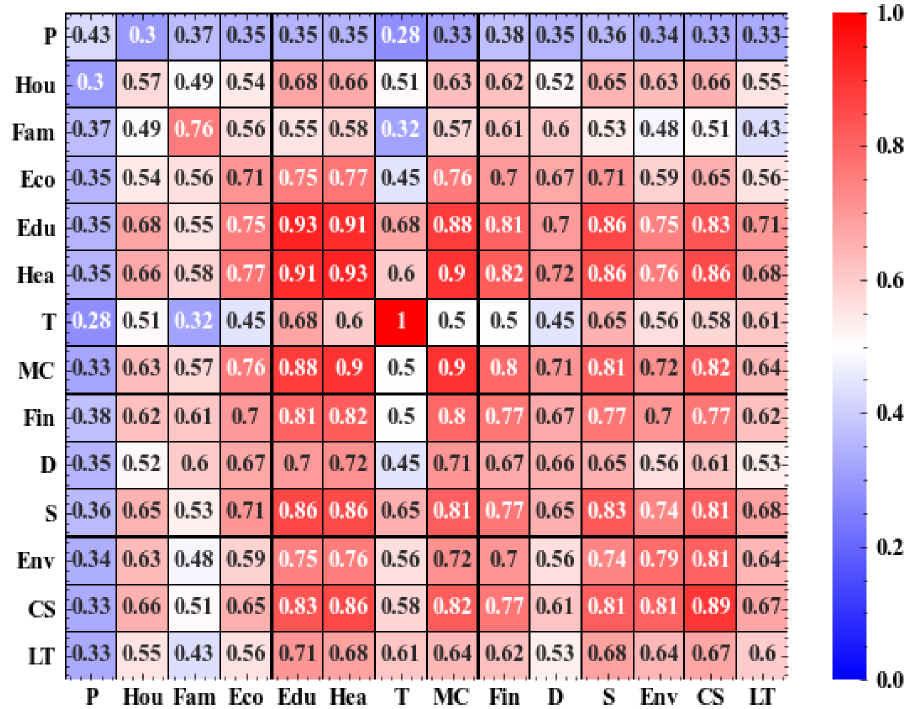
Figure A - 3 Heatmap of correlation coefficients between proxy features of DECP and Acorn option properties with ACC > 0.8. (Only representative parts of results are shown for viewability)



(a) The 75 options with strong correlation to DEC;



(b) The 186 options with strong correlation to DECP;



(c) The 206 options with strong correlation to DEC or DECP.

Figure A - 4 - Heatmap of correlation coefficients between the strongly ECB correlated option properties by mean ACC of major items.

References

[1] IEA. World Energy Outlook 2020. Paris, 2020. https://www.iea.org/reports/world-energy-outlook-2020.

[2] Perera KS, Aung Z, Woon WL. Machine Learning Techniques for Supporting Renewable Energy Generation and Integration: A Survey. Cham: Springer International Publishing; 2014. p. 81-96. 10.1007/978-3-319-13290-7_7.

[3] Dawoud SM, Lin X, Okba MI. Hybrid renewable microgrid optimization techniques: A review. Renewable and Sustainable Energy Reviews. 2018;82:2039-2052. DOI:10.1016/j.rser.2017.08.007.

[4] Schofield JR. Dynamic time-of-use electricity pricing for residential demand response: Design and analysis of the Low Carbon London smart-metering trial: Imperial College London; 2015. DOI:10.13140/RG.2.1.2663.6644/2.

[5] Behl M, Smarra F, Mangharam R. DR-Advisor: A data-driven demand response recommender system. APPL ENERG. 2016;170:30-46. DOI:10.1016/j.apenergy.2016.02.090.

[6] Salcedo-Sanz S, Cornejo-Bueno L, Prieto L, Paredes D, García-Herrera R. Feature selection in machine learning prediction systems for renewable energy applications. Renewable and Sustainable Energy Reviews. 2018;90:728-741. DOI:10.1016/j.rser.2018.04.008.

[7] 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. Renewable and Sustainable Energy Reviews. 2018;82:1027-1047. DOI:10.1016/j.rser.2017.09.108.

[8] Chitalia G, Pipattanasomporn M, Garg V, Rahman S. Robust short-term electrical load forecasting framework for commercial buildings using deep recurrent neural networks. APPL ENERG. 2020;278:115410. DOI:10.1016/j.apenergy.2020.115410.
[9] Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. APPL ENERG. 2019;235:1072-1089. DOI:10.1016/j.apenergy.2018.11.002.
[10] Perera ATD, Wickramasinghe PU, Nik VM, Scartezzini J. Machine learning methods to assist energy system optimization. APPL ENERG. 2019;243:191-205. DOI:10.1016/j.apenergy.2019.03.202.
[11] Piao M, Shon HS, Lee JY, Ryu KH. Subspace Projection Method Based Clustering Analysis in Load Profiling. IEEE T POWER SYST. 2014;29:2628-2635. DOI:10.1109/TPWRS.2014.2309697.
[12] Torriti J. A review of time use models of residential electricity demand. Renewable and Sustainable Energy Reviews. 2014;37:265-272. DOI:10.1016/j.rser.2014.05.034.
[13] Fan C, Xiao F, Li Z, Wang J. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. ENERG BUILDINGS. 2018;159:296-308. DOI:10.1016/j.enbuild.2017.11.008.
[14] Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. ENERGY. 2012;42:68-80. DOI:10.1016/j.energy.2011.12.031.
[15] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. APPL ENERG. 2014;135:461-471. DOI:10.1016/j.apenergy.2014.08.111.
[16] Iglesias F, Kastner W. Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. ENERGIES. 2013;6:579-597. DOI:10.3390/en6020579.
[17] Toit JD, Davimes R, Mohamed A, Patel K, Nye JM. Customer Segmentation Using Unsupervised Learning on Daily Energy Load Profiles. Journal of advances in information technology. 2016;7:69-75. DOI:10.12720/jait.7.2.69-75.
[18] Aghabozorgi S, Shirkhorshidi AS, Wah TY. Time-series clustering – A decade review. INFORM SYST. 2015;53:16-38. DOI:10.1016/j.is.2015.04.007.
[19] Guo Z, Zhou K, Zhang C, Lu X, Chen W, Yang S. Residential electricity consumption behavior: Influencing factors, related theories and intervention strategies. Renewable and Sustainable Energy Reviews. 2018;81:399-412. DOI:10.1016/j.rser.2017.07.046.
[20] Dorfer M, Kelz R, Widmer G. Deep Linear Discriminant Analysis. International Conference on Learning Representations, ICLR. 2016. https://arxiv.org/abs/1511.04707.
[21] Frey BJ, Dueck D. Clustering by passing messages between data points. SCIENCE. 2007;315:972-976. DOI:10.1126/science.1136800.
[22] Markovič R, Gosak M, Grubelnik V, Marhl M, Virtič P. Data-driven classification of residential energy consumption patterns by means of functional connectivity networks. APPL ENERG. 2019;242:506-515. DOI:10.1016/j.apenergy.2019.03.134.
[23] Satre-Meloy A, Diakonova M, Grünewald P. Cluster analysis and prediction of residential peak demand profiles using occupant activity data. APPL ENERG. 2020;260:114246. DOI:10.1016/j.apenergy.2019.114246.
[24] Kwac J, Flora J, Rajagopal R. Household Energy Consumption Segmentation Using Hourly Data. IEEE T SMART GRID. 2014;5:420-430. DOI:10.1109/TSG.2013.2278477.
[25] McLoughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. APPL ENERG. 2015;141:190-199. DOI:10.1016/j.apenergy.2014.12.039.
[26] Jalori S, Reddy TA, "A new clustering method to identify outliers and diurnal schedules from building energy interval data," in 2015 ASHRAE Annual Conference, (ASHRAE Transactions, Atlanta, 2015), pp. 33-44.
[27] Stuhlsatz A, Lippel J, Zielke T. Feature extraction with deep neural networks by a generalized discriminant analysis. IEEE Transactions on Neural Networks & Learning Systems. 2012;23:596-608. DOI:10.1109/TNNLS.2012.2183645.
[28] Liao TW. Clustering of time series data — a survey. PATTERN RECOGN. 2005;38:1857-1874. DOI:10.1016/j.patcog.2005.01.025.
[29] Low Carbon London project: Data from the dynamic time-of-use electricity pricing trial., 2013. https://innovation.ukpowernetworks.co.uk/projects/low-carbon-london/.
[30] CACI. Acorn consumer classification data., 2013. http://acorn.caci.co.uk.
[31] Carmichael R, Schofield J, Woolf M, Bilton M, Ozaki R, Strbac G. Residential consumer attitudes to time-varying pricing, Report A2 for the Low Carbon London LCNF project.: Imperial College London; 2014. https://innovation.ukpowernetworks.co.uk/wp-content/uploads/2019/05/A2-Residential-Consumer-Attitudes-to-Time-varying-Pricing.pdf.
[32] Schofield J, Carmichael R, Tindemans S, Woolf M, Bilton M, Strbac G. Residential consumer responsiveness to time-varying pricing, Report A3 for the Low Carbon London LCNF project.: Imperial College London; 2014. https://innovation.ukpowernetworks.co.uk/wp-content/uploads/2019/05/A3-Residential-Consumer-Responsiveness-to-Time-varying-Pricing.pdf.

[33] Hodges JL. The significance probability of the smirnov two-sample test. Arkiv Fr Matematik. 1958;3:469-486. DOI:10.1007/BF02589501.
[34] Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. Annals of eugenics. 1936;7:179-188. DOI:10.1111/j.1469-1809.1936.tb02137.x.
[35] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J MACH LEARN RES. 2011;12:2825-2830. DOI:10.5555/1953048.2078195.
[36] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2016. https://arxiv.org/abs//1603.04467.
[37] Bergstra J, Yamins D, Cox D, "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures," in Proceedings of the 30th International Conference on Machine Learning, PMLR, (2013), pp. 115-123.
[38] Torriti J, Yunusov T. It’s only a matter of time: Flexibility, activities and time of use tariffs in the United Kingdom. ENERGY RES SOC SCI. 2020;69:101697. DOI:10.1016/j.erss.2020.101697.
[39] 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;24:891-912. DOI:10.1177/1420326X15600776.
[40] Ramírez-Mendiola JL, Grünewald P, Eyre N. Residential activity pattern modelling through stochastic chains of variable memory length. APPL ENERG. 2019;237:417-430. DOI:10.1016/j.apenergy.2019.01.019.
[41] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews. 2018;81:1192-1205. DOI:10.1016/j.rser.2017.04.095.
[42] Hubert L, Arabie P. Comparing Partitions. J CLASSIF. 1985;2:193-218. DOI:10.1007/BF01908075.
[43] Vinh NX, Epps J, Bailey J, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" in Proceedings of the 26th Annual International Conference on Machine Learning, ICML, (ACM, 2009), pp. 1073-1080. DOI:10.1145/1553374.1553511.
[44] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational & Applied Mathematics. 1987;20:53-65. DOI:10.1016/0377-0427(87)90125-7.
[45] Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell. 1979;PAMI-1:224-227. DOI:10.1109/TPAMI.1979.4766909.