Classification and characterization of intra-day load curves of PV and non-PV households using interpretable feature extraction and feature-based clustering

Maomao Hu[a,1], Dongjiao Ge[a], Rory Telford[b], Bruce Stephen[b] and David C. H. Wallom[a]

[a] Oxford e-Research Center, Department of Engineering Science, University of Oxford, Oxford OX1 3QG, United Kingdom

[b] Institute for Energy and Environment, Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1RD, United Kingdom

Abstract

Load pattern categorization plays a significant role in enhancing the understanding of demand characteristics of different cohorts of energy customers on a distribution network. For a distribution network integrating with photovoltaics (PV) systems, it is also desirable to identify PV households since embedded unauthorized PVs have detrimental impacts on distribution networks regarding voltage control, frequency regulation, and back-feeding flow. We developed a holistic smart meter data analytics approach to classifying and characterizing the intra-day load curves of PV and non-PV households. Unlike existing studies based on raw time-series data, a series of interpretable and discriminating global and peak-period features were first developed to extract the physical information of load patterns. A two-step feature-based clustering was then applied to classify the load profiles of PV and non-PV households. A post-clustering quantitative compositional analysis was developed to characterize the energy use variability and compositional changes for each household. The proposed PV household identification method allows electricity

[1] Corresponding author.

E-mail address: maomao.hu@eng.ox.ac.uk

suppliers to identify the existence of PV installations in a particular area for voltage control and ancillary service provision. Besides, effective load pattern categorization and post-clustering characterization can help better understand demand characteristics of different cohorts of customers, network utilization, and network-level changes in load growth.

Keywords: feature extraction; feature-based clustering; compositional data analysis; smart meter data analytics; unsupervised machine learning.

1. Introduction

Today's electrical grids are facing power imbalance and network capacity issues, caused by vehicle electrification and increasing penetration of intermittent renewable energy sources (RESs) [1-3]. To relieve these issues, demand-side management has been proposed and demonstrated as one of the effective solutions deployed at both individual customer premise level [4-6] and distribution network level [7]. Enhanced understanding of electricity load patterns of end-users can help improve demand-side management performance and load forecasting accuracy [8, 9]. Additional benefits include helping electricity suppliers to formulate feasible real-time pricing structures for different cohorts of customers with different energy use behavior and lifestyles [10, 11]. As of September 2020, 39% of domestic gas meters and 43% of domestic electricity meters were smart meters in the UK [12]. Increasing deployment of smart meters has made high-granularity energy consumption data available, providing the opportunity to better understand the electricity load patterns of residential end-users. How to extract actionable information and knowledge from large volumes of smart meter data to address practical issues is increasingly recognized as a significant and challenging research topic [13]. In this study, the smart meter data were used to fulfill two tasks: *1) PV household identification* and *2) load pattern categorization and characterization*.

1.1. Identification of domestic PV installation

With the proliferation of distributed PV installations on distribution networks, it is key for utility companies to gain visibility of all solar prosumers. However, utility companies normally do not have locations of all installed PV systems due to unregistered customers. Reasons for unregistered installations include fee avoidance, lack of awareness, change in ownership, and mistaken data entry [14]. The lack of accurate understanding of PV installations has detrimental effects on the operation of distribution networks in terms of voltage regulation, frequency control, circuit reconfiguration, and back-feeding flow [14, 15]. Effective identification of all PV households on a distribution network is therefore needed to avoid these hidden issues.

PV household identification/detection has received increasing attention, and a few attempts have been made to this topic using a wide variety of methodologies. Malof et al. [16] developed an algorithm to automatically detect PV panel locations based on high-resolution color aerial imagery. The algorithm was tested on aerial imagery (135 km^2) over a city in California, USA. Results demonstrated that the proposed algorithm could effectively identify the locations of PV panels, but lacked the ability to accurately estimate the size of the PV panel. Zhang and Grijalva [14] used a change-point detection algorithm to identify anomalous energy use profiles. A statistical inference approach was then adopted to determine whether a given change point was caused by the installation of PV panels. Compared with [16], after the PV location identification, the size (i.e., rated power) of the PV panel was further estimated with the assistance of the local cloud cover index. However, the energy use data before and after the installation of the PV system was required to detect the change point, which was difficult to obtain. In [17], Wang et al. proposed a two-stage PV size estimation approach based on support vector (SV) machine algorithm and weather features, including SV classification-based PV generation detection and SV regression-

based PV size estimation. Datasets of 183 domestic households with PV generation in Texas, USA were used to test the proposed approach. Results showed that the proposed two-stage approach was able to accurately estimate the size of the PV system.

For the discussed existing studies on PV household identification, various exogenous data are required for each approach, including high-resolution aerial imagery [16], energy use data before and after PV installation [14], and weather data [17]. However, these accurate exogenous data are not always readily available. With the availability of high-granularity smart meter data, it is valuable to investigate the use of a smart meter data analytics approach to identifying PV households on a distribution network.

1.2. Load pattern categorization

Conventionally, energy customers are categorized by building function (e.g., residential, commercial, and industrial customers) or Energy Usage Intensity (EUI) (e.g., customers with high/low EUI). With the availability of high-resolution smart meter data, it is possible to further categorize customers by the similarity of energy use patterns. Load pattern categorization[2], therefore, has become a commonly-used and fundamental task associated with smart meter data analytics [13, 18, 19]. It aims to use suitable clustering techniques to undertake a natural partitioning of load patterns for enhancing understanding of demand characteristics of different groups of energy use customers on a distribution network.

A considerable amount of literature has been published on different clustering techniques for load pattern categorization, including K-means [20-22], hierarchical clustering [23, 24], self-organizing

[2] Note that some other terms carrying the same meaning of "categorization" have also been used in related studies such as "grouping", "classification", "segmentation", "profiling", and "sub-profiling".

map (SOM) [25, 26], and Dirichlet process mixture model (DPMM) [27]. Among these clustering techniques, K-means and hierarchical clustering are two of the most common clustering methods. The K-means was also used in this study for feature-based clustering. Carmo and Christesen [20] applied k-means to group the daily heating load profiles of 139 Danish households with heat pumps. Two main clusters were identified for the heating load patterns on weekdays and at weekends, respectively. The differences between the two clusters were a result of the differences in building characteristics and space heating systems. Gianniou et al. [21] also adopted k-means to segment the daily load patterns of around 8,000 district heating customers in Denmark. 5 types of clusters were classified based primarily on the intensity of energy consumption. Correlation analysis based on the clustering results showed that old buildings and buildings with large floor sizes consumed more energy and were mostly located within the cluster of high energy use intensity. Compared with building age and floor size, the household size had a smaller effect on the cluster membership of the household. Zhan and Liu et al. [22] applied k-means clustering to the categorization of the load profiles of 81 buildings in Singapore. Comparative results showed that the developed clustering-based framework significantly improved the energy benchmarking performance. In [23], hierarchical clustering was used to segment the load patterns of 2000 Irish households and then applied to intra-day energy use prediction. Results indicated that the proposed clustering-based method was able to satisfactorily improve the accuracy of intra-day load forecasting. Jota and Silva et al. [24] adopted hierarchical clustering to identify the typical patterns in the load curves of a hospital building. Results showed that the clustering could help to effectively and quickly predict the building energy consumption and peak demands.

In some studies, multiple clustering methods previously described were used for load curve categorization [28-30]. Specifically, Chicco and Napoli et al. [28] applied k-means, fuzzy k-means,

hierarchical clustering, SOM, and modified follow-the-leader clustering methods to segment the electricity load profiles of 234 non-domestic buildings. The comparative results indicated that the hierarchical clustering and modified follow-the-leader outperformed other clustering techniques. In order to evaluate the effects of temporal resolutions on clustering results, Granell et al. [29, 31] adopted three clustering techniques, including k-means, hierarchical clustering, and DPMM, to group the daily load curves of 197 Bulgarian and British households with various temporal resolutions (0.5 min - 240 min). Results indicated that to efficiently segment the electricity end-users for most applications, the temporal resolution of load profiles needed to be at least 30-min and ideally 8 or 15 min.

Most studies on load pattern categorization have focused on using raw time-series data based on unsupervised machine learning algorithms. However, as pointed out by Timmer [32], "*the crucial problem is not the classificator function (linear or nonlinear), but the selection of well-discriminating features. In addition, the features should contribute to an understanding of the properties of* …." Similarly, clustering methods based on interpretable and discriminating features are also desirable for load pattern categorization, which can provide a better understanding of the energy use behavior of customers and changes in load growth at the distribution network level.

1.3. Innovations and contributions

In summary, for a low-voltage distribution network with an unknown level of PV penetration, it is desirable to use smart meter data analytics to fulfill the tasks of PV household identification and load pattern categorization, which can help enhance the understanding of the magnitude, heterogeneity, and diversity of the electricity load profiles at a distribution network level. Therefore, this paper describes a holistic data analytics approach to classifying and characterizing the intra-day load curves of PV and non-PV households using interpretable feature extraction,

feature-based clustering, and post-clustering quantitative compositional analysis. The major innovations and contributions of this study are:

1) A series of interpretable features, including global features and peak-period features, are developed to extract the physical information in daily load patterns. A symbolic representation technique is introduced to transform the raw time-series patterns into alphabetical words to help with the motif discovery and peak-period feature extraction.
2) A two-step feature-based clustering, i.e., monthly and yearly feature-based clustering, is proposed to classify and characterize the electricity load profiles of PV and non-PV households. The monthly clustering analysis is first carried out to identify the presence of PV households for further yearly high-granularity clustering. Before the feature-based clustering, dimensionality reduction is performed to select the top discriminating features for computational load reduction.
3) In order to quantitatively measure the effects of the presence or absence of PV systems on domestic energy consumption behavior, the post-clustering quantitative compositional analysis is developed to characterize the energy consumption variability and compositional change over a week/year for each household. The compositional change is based on Aitchison distance, which helps to quantify the changes in cluster composition for each household.

The rest of this paper is organized as follows: Section 2 introduces the overview of the proposed holistic data analytics method. Section 3 presents the whole methodology, including the data preprocessing method (Section 3.1), the development of the interpretable global and peak-period features (Section 3.2), and a two-step feature-based clustering method (Section 3.3). In Section 4, the proposed method is tested based on a representative load pattern dataset with mixed types of households. Finally, conclusions are presented in Section 5.

2. Overview of the proposed approach

As shown in Fig. 1, the developed holistic smart meter data analytics approach consists of four stages. In Stage 1, the raw time-series smart meter data are first cleaned up and transformed into alphabetical words using the symbolic representation method. The symbolic words are then used for motif discovery. In Stage 2, we develop and extract a number of interpretable features for the motif patterns obtained from Stage 1. The interpretable features for each daily load profile include global features and peak-period features. A two-step feature-based clustering is then carried out in Stage 3 and Stage 4. The objective of the monthly feature-based clustering in Stage 3 is to classify the PV and non-PV households by analysing the cluster membership. In Stage 4, yearly clustering and post-clustering quantitative compositional analyses are carried out to characterize the energy use variability and compositional change of the load profiles for PV and non-PV households, respectively. In both Stage 3 and Stage 4, the dimension of the features is reduced by using principal component analysis (PCA) before the k-means based clustering.

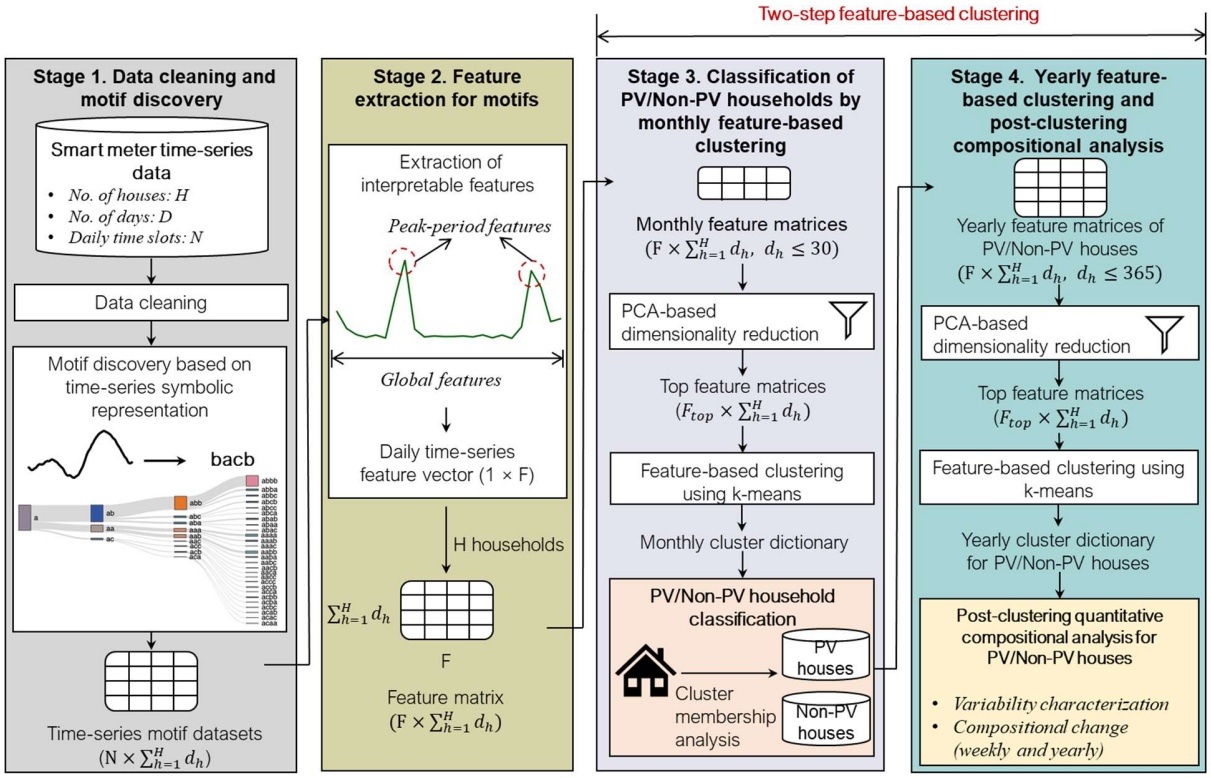

Fig. 1. Flowchart of the proposed approach.

3. Methodology

3.1. Data cleaning and time-series motif discovery

Effective data preprocessing can ensure the analysed load patterns are representative of the co-occurring and temporal aspects of load behaviour captured via the high-resolution meter readings. Data preprocessing therefore is first performed in Stage 1, including data cleaning and time-series motif discovery (i.e., data transformation). Missing values are a common issue in the electricity load data collected from smart meters. As such data cleaning and imputation are required to improve the quality of the raw data by filling the missing values. Details of the data cleaning and imputation will be elaborated in the case study.

To quickly identify the motif patterns within the smart meter data, the technique of Symbolic Aggregate approXimation (SAX) representation [33, 34] is used in this study to transform a raw time series of length L_{raw} to a string of length L_{sax} ($L_{sax} < L_{raw}$). The whole process of SAX representation involves 3 distinct steps: Z-score normalization, segmentation, and discrete representation. Given $x(t)$ is a univariate daily load time series with the length of L_{raw}, $x(t)$ is first standardized by using Z-score normalization, i.e., $Z(t) = \frac{x(t)-\mu}{\sigma}$. For a group of households, the mean value μ and standard deviation σ are calculated based on the load profiles of all households, rather than calculated individually. The normalized time series $Z(t)$ is then divided into N_{sub} sub-sequences with equal lengths to reduce the dimensionality. The mean value of the data within a sub-sequence is computed and used to represent this subsequence. This segmentation representation is also known as Piecewise Aggregate Approximation (PAA) [35]. Last, the PAA representation of a time series is discretized and mapped into alphabetic letters. The alphabetic letter is determined by where the mean value falls within a series of breakpoints. Given that the normalized time series has Gaussian distribution, the breakpoints split the Gaussian distribution into a number of equiprobable regions. The alphabet size A equals the number of the equally sized areas under the Gaussian curve. The breakpoints for the value of A ranging from 3 to 10 can be looked up in [33].

The results of SAX representation majorly depend on the choices of sub-sequence number N_{sub} ($L_{sax} = N_{sub}$) and the alphabet size A. Larger values of N_{sub} and A result in more SAX patterns. For the motif discovery of intra-day load profiles, the settings of $N_{sub} = 4$ and $A = 3$ are recommended to balance the number of generated SAX patterns and resolution [36]. An example of the whole SAX representation procedure is given in Fig. 2. A power consumption time series with 48 points is converted into a 4-letter SAX word. The SAX word can help simply interpret the

daily load profile from a physical perspective. For example, the letter *a* for 00:00 - 06:00 and the letter *c* for 18:00 - 24:00 represent the low-level and high-level energy consumption, respectively.

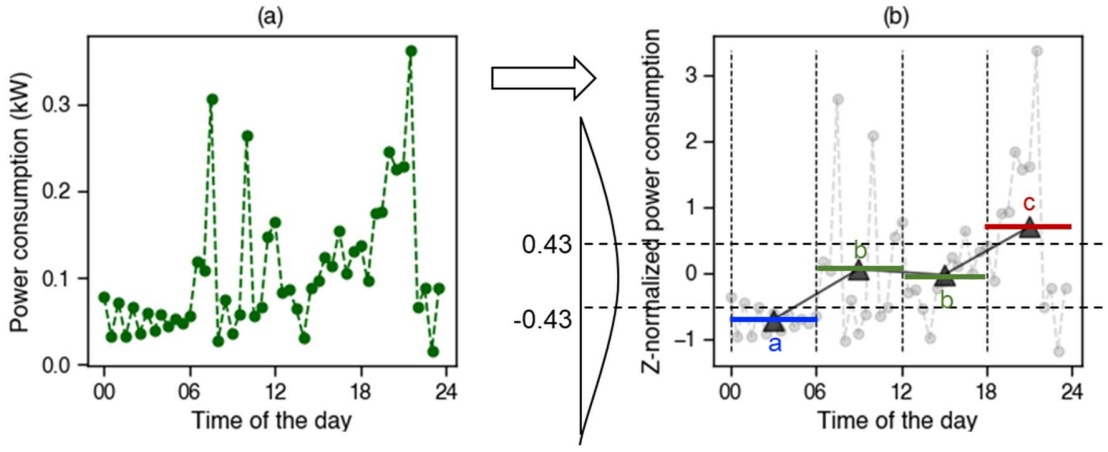

Fig. 2. An example of the SAX representation of a daily load pattern: a) a 30-min daily load pattern; b) Z-normalized pattern and SAX word. (Note: the length of raw time series L_{raw}= 48; the sub-sequence number and the length of SAX string $N_{sub} = L_{sax}$ = 4; alphabet size A = 3; the mapped SAX word is *abbc*)

The evolution of the daily load profiles can be clearly visualized by using an augmented suffix tree of SAX strings and substrings with the assistance of a Sankey diagram [36], which will be elaborated in the case study. Besides being used for motif detection, the SAX representation will also be used to capture the dynamics of the daily load pattern in Section 3.2.

3.2. Interpretable feature extraction

Feature-based clustering is proposed in this study to contribute to the understanding of the dynamics of load patterns. Before the feature-based clustering in Section 3.3, a set of interpretable features (21 in total) are first developed and extracted in this section. As shown in Table 1, the 21 features can be categorized into 13 global features (GFs) and 8 peak-period features (PFs). All 13 GFs are extracted based on raw time-series data, while the 8 PFs are extracted based on SAX

words. For better illustration, Fig. 3 shows the graphical representation of some global and peak-period features of a typical load pattern.

Table 1. 21 interpretable features for daily electricity load pattern.

Feature No		Physical meaning of each feature	Feature category	Data type for extraction
Global features	GF-01	Mean value of a daily load pattern	Statistical	Raw time-series data
	GF-02	Stand deviation of a daily load pattern		
	GF-03	Maximum power consumption during a day		
	GF-04	Minimum power consumption during a day		
	GF-05	Range of power consumption during a day (i.e., maximum - minimum)		
	GF-06	Percentage of values above mean value		
	GF-07	Sum of net loads during business hours (9:00 – 18:00)		
	GF-08	Sum of net loads during non-business hours		
	GF-09	Skewness of the distribution of a daily load pattern		
	GF-10	Kurtosis of the distribution of a daily load pattern		
	GF-11	Mode of the 5-bin histogram for a daily load pattern		
	GF-12	Longest subsequence where consecutive value above mean value	Temporal	
	GF-13	Longest period of successive increase		
Peak-period features	PF-01	Number of peak periods		Symbolic SAX words
	PF-02	Occurrence time (starting time) of each peak period		
	PF-03	Shortest time interval between peaks if more than one peak exists		
	PF-04	Duration of each peak		
	PF-05	Occurrence time of the longest peak period		
	PF-06	Duration of the longest peak period		
	PF-07	Upward slope of the longest peak		
	PF-08	Downward slope of the longest peak		
Note: Net load is obtained by subtracting the behind-the-meter PV generation from the actual load.				

Among the global features, the first 11 features from GF-01 to GF-11 are statistical features. Since the features of GF-01 to GF-06 are basic statistical metrics, their physical meanings are not elaborated here. The GF-07 and GF-08 features denote the sums of net household loads during business hours (9:00 – 18:00) and non-business hours (18:00 – 8:00), respectively. Skewness, i.e., GF-09, is a measure of the asymmetry of the distribution of a daily load profile about its mean value. For a daily load pattern, positive skewness indicates data are right-tailed and its mean value is skewed to the right of the daily median value. Kurtosis, i.e., GF-10, is used to measure the tail-heaviness of the distribution of a daily load profile, i.e., whether the distribution is heavy-tailed or light-tailed relative to a normal distribution. If the kurtosis is negative, the distribution is light-tailed; otherwise, the distribution is heavy-tailed. For a daily load profile with some extreme values, it might have extremely heavy tails (i.e., a large value of kurtosis). The feature of GF-11 (Mode of the 5-bin histogram) refers to the data value where the histogram with 5 bins reaches its peak. Unlike the first 11 global features, the remaining 2 global features, i.e., GF-12 and GF-13, are temporal features, which are used to characterize the temporal properties of the time-series load pattern.

8 temporal features for peak periods are also developed in this study. Besides being used for motif discovery, the symbolic SAX words are also used to help quickly extract the temporal features for peak periods. To characterize the daily load pattern a higher resolution, the number of subsequence (N_{sub}) and the alphabet size (A) are set as 24 and 7, respectively, as shown in Fig. 3-b. After being transformed into alphabet letters, several peak-period features can be quickly identified by searching the SAX letter of "g", including peak number (PF-01), occurrence time of peak (PF-02 and PF-05), peak duration (PF-04 and PF-06), and the shortest interval between two peaks (PF-03). In addition to those simple features, the upward and downward slopes of a specific peak can

also be recognized by analyzing the change of the SAX letter level (first discrete difference) at each time step.

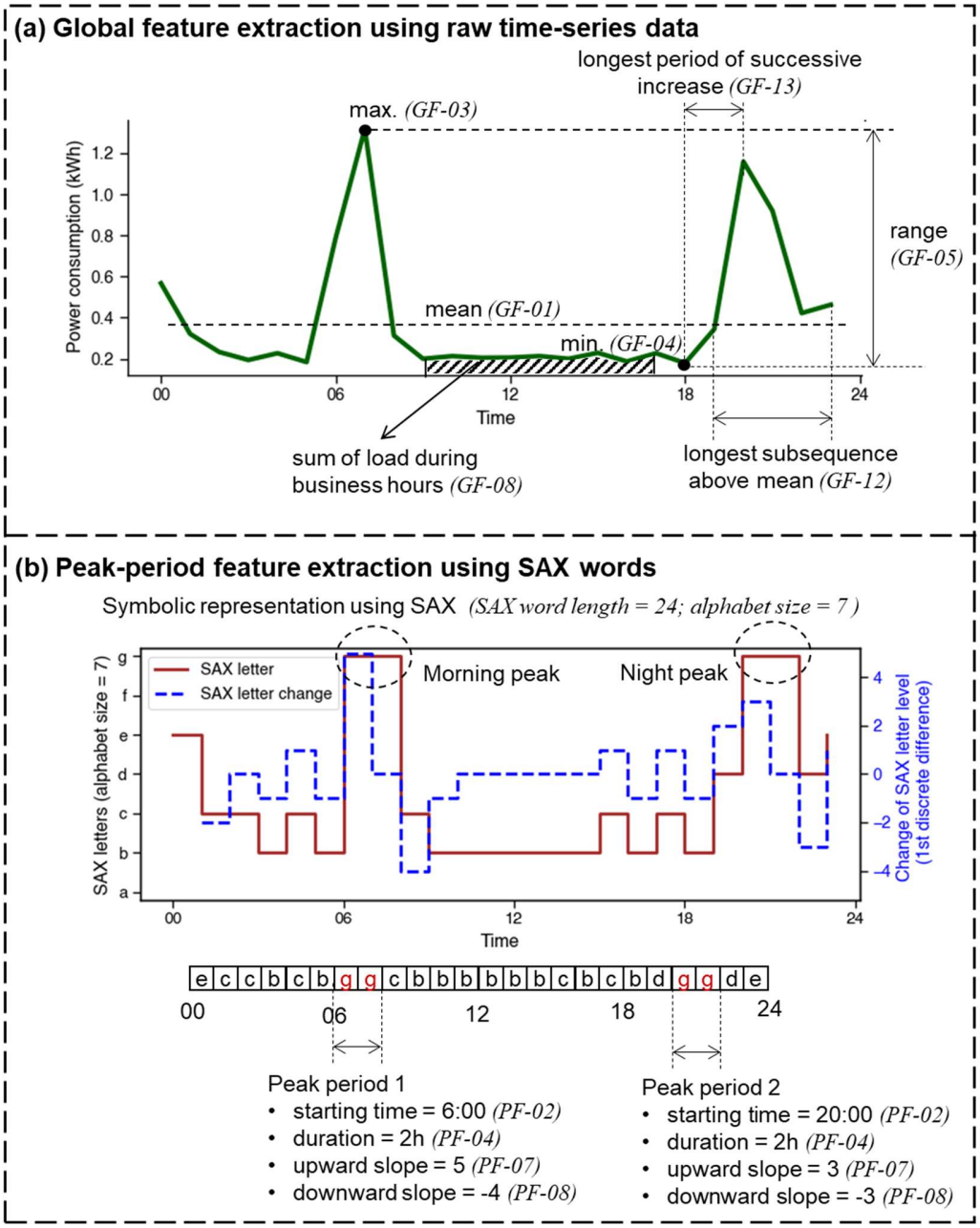

Fig. 3. Graphical representation of the global and peak-period interpretable features of a typical load pattern: a) Global feature extraction; b) Peak-period feature extraction.

An example of the peak-period feature extraction for a typical load pattern is given in Fig. 3-b. Based on the SAX representation of the load profile, two peaks represented by the letter "g" can be easily located: a two-hour morning peak with starting time slot of 6:00 and a two-hour nighttime peak with starting time slot of 20:00. Then, the change of SAX letter level at each time step can be derived based on the SAX words. The upward slope and downward slope can be recognized by analyzing the change of SAX letter level. For the morning peak, the upward and downward slopes are 5 and -4, respectively. For the nighttime peak, the upward and downward slopes are 3 and -3, respectively. Compared with the night peak, the morning peak with steeper slopes, therefore, has worse effects on the reliability of the electrical grids for the specific daily load pattern.

The whole process of interpretable feature extraction was performed in Python 3. We have packaged our code into a package and distributed using both GitHub and PyPI [37]. The interpretable features developed and extracted in this section are general features. Prior to the data mining process, dimensionality reduction is required to select the top discriminating features and to reduce the computational loads, which will be introduced in Section 3.3.

3.3. A two-step feature-based clustering and post-clustering compositional analysis

In this section, a two-step feature-based clustering method is proposed to segment the daily load profiles, as shown in Fig. 1. In both steps, dimensionality reduction is implemented in advance to reduce the dimension of the feature matrices and to reduce the runtime of the clustering algorithm.

3.3.1. PCA-based dimensionality reduction

PCA is a widely used statistical tool to reduce the dimensionality of a large multivariate dataset, but at the same time minimizing the information loss and retaining as much variability as possible [38, 39]. The task of PCA is to find a set of dominant and orthogonal variables (i.e., principal

components) to represent the most important information contained by original inter-correlated variables. The criterion of the determination of the number of principal components (k) can be found in [40].

As shown in Fig. 1, the task of dimensionality reduction in this study is to convert the dimension of the matrix from F $\times \sum_{h=1}^{H} d_h$ to $F_{top} \times \sum_{h=1}^{H} d_h$ using PCA, where d_h is the number of days of household h after data cleaning and motif discovery; H denotes the total number of households; F denotes the feature number of high-dimensional representation; and F_{top} denotes the reduced number of the top features in a low-dimensional space ($F_{top} \ll F$).

3.3.2. Feature-based clustering using k-means

The clustering analysis used in this study is k-means, which is an iterative algorithm segmenting the dataset into K clusters with the objective of minimization of the within-cluster sum of all distances (Euclidean distance) to their corresponding cluster centers. Silhouette score [41] and Calinski-Harabasz score [42] are used in this study to measure the clustering performances. Silhouette score is the average Silhouette coefficient, as defined in Eq. (1), of all points and the Silhouette coefficient of each point indicates how close each point is to its own cluster compared to other neighboring clusters, ranging from [-1, 1]. The Calinski-Harabasz score is used to quantify the ratio between the within-cluster dispersion and the between-cluster dispersion, as defined in Eqs. (2) – (4). The selection of K is a trade-off between of segmentation resolution (i.e., cluster number) and accuracy of representativeness. It is expected to classify samples into more tightly grouped and well-separated clusters. The commonly used 'elbow' method is adopted in this study to tune the selection of K [22, 36].

$$S_{sil}(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}} \quad (1)$$

where $a(i)$ denotes the average distance between a sample i and all other points in the same cluster and $b(i)$ denotes the average distance between a sample i and all other points in the next nearest cluster.

$$S_{cal} = \frac{tr(B_k)}{tr(W_k)} \times \frac{N_A - k}{k-1} \tag{2}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{3}$$

$$B_k = \sum_{q=1}^{k} N_q (c_q - c_A)(c_q - c_A)^T \tag{4}$$

where A is a set of points with the size of N_A and the center of c_A, and it is clustered into k clusters; $tr(B_k)$ and $tr(W_k)$ denote the trace of the between group dispersion matrix and the trace of the within-cluster dispersion matrix, respectively; and C_q is the set of points in cluster q with the size of N_q and the center of c_q.

3.3.3. A two-step feature-based clustering

As shown in Fig. 1, the whole procedure includes two parts: monthly feature-based clustering for PV/non-PV household classification and yearly feature-based clustering for each category of households.

- *Monthly feature-based clustering for classification*

To improve the granularity of clustering results, the first step is to separate the PV-houses from the mixed households. If summer month data is available, a month-long dataset is adequate for efficient classification of PV and non-PV households, i.e., $d_h \leq 30$. For other months, longer time series may be required to give adequate results. Summer months are recommended since the PV houses normally generate a considerable amount of solar energy in summer due to high irradiance,

which helps to clearly distinguish the PV houses from non-PV houses in the cluster membership analysis.

After the dimensionality reduction and feature-based clustering, the monthly cluster dictionary C_{month} for all types of houses can be obtained, which includes cluster centroids above the y-axis ($C_{month}^{positive}$) and below the y-axis ($C_{monthly}^{negative}$). For a centroid in $C_{month}^{negative}$, its negative part is simply caused by the negative value of the net load, which is obtained by subtracting the behind-the-meter PV generation from the actual load. For the household i, the cluster membership of its load patterns, $C_{month\ ,i}$, is a subsequence of $C_{monthly}$. If $C_{monthly,i}$ contains any element in $C_{monthly}^{negative}$, then household i is regarded as a household with PVs; otherwise, it is a household without PVs. In this way, the mixed households can be efficiently classified into PV households and non-PV households for high-granularity yearly clustering.

- *Yearly feature-based clustering and post-clustering compositional analysis*

After the classification, the yearly feature-based clustering analysis is carried out to further segment the patterns within PV households and non-PV households, respectively. The yearly cluster dictionaries of PV households (C_{yearly}^{PV}) and non-PV households (C_{yearly}^{non-P}) can be obtained by using the previously described PCA-based feature reduction and k-means clustering technique. Moreover, the post-clustering quantitative compositional analysis is proposed in this study to characterize the energy use variability and compositional change of each household. This allows us to understand how the presence or absence of PV systems change domestic energy consumption behavior and to measure any changes quantitatively.

Variability characterization: Based on the cluster dictionaries, the metric of entropy is applied to quantify the variability of energy use. For the household i, given its cluster membership $C_{yearly,i}$ is $[c^1_{yearly,i}, c^2_{yearly,i}, \dots, c^n_{yearly,i}, \dots, c^K_{yearly,i}]$, its entropy, En_i, can then be obtained by Eq. (5).

$$En_i = -\sum_{n=1}^{K} p\left(c^n_{yearly,i}\right) \times \log p(c^n_{yearly,i}) \tag{5}$$

where $p\left(c^n_{yearly,i}\right)$ denotes the percentage of Household i's daily load profiles within cluster $c^n_{yearly,i}$ across an annual period. The entropy reaches the peak when all cluster memberships are included and with the same occurrence possibility, i.e., $p\left(c^n_{yearly,i}\right) = 1/K$. The smallest entropy (i.e., $En_i = 0$) happens when only one cluster centroid exists (i.e., $K = 1$). The higher entropy of a household indicates the higher variability and uncertainty of the energy use behavior.

Compositional change*:* In addition to the characterization of energy use variability, the method of compositional data analysis is applied in this study to characterize the change in cluster composition for each household. Compositional data measures each sample as a set of proportions, which is a vector of non-negative values containing relative information [43]. For compositional data, the sample space is the simplex, S^D, defined in Eq. (6).

$$S^D = \left\{\boldsymbol{x} = [x_1, x_2, \dots, x_l, \dots, x_D] | x_l \geq 0 \; and \; \sum_{j=1}^{D} x_j = c\right\} \tag{6}$$

where D denotes the number of variables of the composition $\boldsymbol{x}$; c depends on the units of the measurements, e.g., 1 for proportions and 100 for percentages (%). In the simplex, S^D, the conventional Euclidean distance is not adequate to quantify the difference between two compositional samples since it doesn't have the properties of scale invariance, permutation invariance, and sub-compositional coherence [44]. The Aitchison distance based on a log-ratio transformation of a composition is therefore used to measure the distance between two D-

dimensional compositions, $\boldsymbol{x}$ and $\boldsymbol{y}$, as defined in Eq. (7) [43, 45]. For the compositional data with zeros, Eq. (7) is not suitable for calculating the Aitchison distance. Accordingly, Eq. (8) is used to replace the original $\boldsymbol{x} \in S^D$ containing zeros with $\boldsymbol{z} \in S^D$ without zeros [46].

$$d_{Aitchison}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{2D}\sum_{l=1}^{D}\sum_{j=1}^{D}(\ln\frac{x_l}{x_j} - \ln\frac{y_l}{y_j})^2} \tag{7}$$

$$z_l = \begin{cases} \delta_l, & if\ x_l = 0 \\ \left(1 - \frac{\sum_{k|x_k=0}\delta_k}{c}\right)x_l, & if\ x_l > 0 \end{cases} \tag{8}$$

where δ_l is the imputed value on x_l; c denotes the constant value of the sum constraint in Eq. (6). For the determination of δ_l, when the compositional dataset contains a small number of zeros, the suggested range is [0.001, 0.01]; otherwise, the recommended range of δ_l is [0.0001, 0.001] [46]. In this study, the changes in cluster composition for each household are investigated at two levels: weekly compositional change over day of the week and yearly compositional change over month of the year.

4. Case study: classification of smart meter load profiles in the UK

4.1. Description of dataset

The dataset used in this study was based on the Energy Demand Research Project (EDRP) [47], which was carried out across Great Britain between 2007 and 2010. It aimed to explore the energy consumption behavior of different British households based on the half-hourly electricity consumption data of around 14,000 smart meters. For non-PV households, the electrical end-use energy of 614 households between 1 July 2008 to 30 June 2009 was extracted directly from the EDRP dataset for use within this case study.

No households that participated in EDRP had PV generation installed; it was, therefore, necessary to simulate net load at households with PV. This simulation process primarily extracted end-use demand from households within the EDRP dataset. Half-hourly PV generation at each household was then simulated by using an open-source PV model [48] combined with global horizontal irradiance (GHI) weather data [49]. Half-hourly GHI data at the specified location of a household is inferred from the Typical Meteorological Year (TMY) data using interpolation among the four closest latitude-longitude grid points. The power rating of the installed PV and GHI data are input to the open-source model to simulate half-hourly PV output. The open-source model has extensive databases that include parameters on different rated PV panels and inverters. The model also accounts for household variables, including roof-pitch, area, and angle to the north. Within this case study, 154 houses were assumed to have PV installed. Roof variables were randomly sampled to model sufficient distribution of household characteristics. Installed PV capacity for each household was related to the roof variables, with the number of PV panels on each house dependent on the respective roof areas, Accordingly, the mean installed capacity across all modeled houses was 3.075kW with a standard deviation of 1.243kW. Note that net loads of PV households are obtained by subtracting the behind-the-meter simulated PV generation from the actual end-use demand extracted from the EDRP dataset.

In summary, the dataset consists of half-hourly electricity net load data of 768 British households, including 614 non-PV households and 154 PV households, between 1 July 2008 to 30 June 2009. The penetration rate of the PV installation is 20%. The daily load profiles of PV and non-PV households are randomly mixed to test whether the monthly feature-based clustering can identify the presence of PV households.

4.2. Data cleaning and time-series motif discovery

- Data cleaning and imputation

One-year (365 days) data are cut off from the original dataset with the principle of removing those months with many missing values. Only those households that contain more than 350 days of completed data are kept. Regarding data imputation, it is well known that similar electricity consumption behaviour can be detected on the same day of the week due to the consistent working routine of each household. For this reason, we compute the median values of the daily usage profile of each day of the week for each household. Specifically, all the daily usage profiles on Mondays of Household *i* are collected; and then compute the median curve of them. Repeating this process for each household from Monday to Sunday and obtaining 7 median usage profiles for each household. These median profiles are further applied to impute missing values of the corresponding household.

- Motif discovery

In the case study, the length of SAX string (L_{sax}) and the alphabet size (A) are set as 4 and 3, respectively. A daily power consumption pattern at 30-minute intervals, i.e., $L_{raw} = 48$, can be then converted into a 4-letter SAX word. After transforming all daily load profiles, the distribution of SAX word patterns can be obtained as shown in Fig. 4. The most frequent SAX word is "*bbbb*", accounting for 46.08% of all load profiles. The time-series motifs and discords can then be separated by setting a threshold of 1%. If the percentage of a SAX word pattern is larger than the threshold, the load pattern is regarded as a motif; otherwise, it is regarded as a discord. In the case study, the motif SAX word patterns include the top 11 most frequent word patterns, representing

95.87% of all electricity load profiles. The rest are regarded as discord patterns and discarded. Note that the threshold can be adjusted to obtain a feasible percentage of discord patterns.

The evolution of the daily load profiles was also visualized by using an augmented suffix tree with the assistance of the Sankey diagram [36] as shown in Fig. 5. The Sankey diagram enables the augmented suffix tree to effectively visualize the occurrence frequencies of all SAX strings and substrings at each level. The height of each substring bar represents the number of substring patterns. For all 768 households, the pattern of "*bbbb*" is found to be the most common pattern, representing the most frequent energy use behavior of customers. Motif patterns except "*cccc*" start with "*b*", indicating a medium level of net load during the period of 00:00 – 06:00.

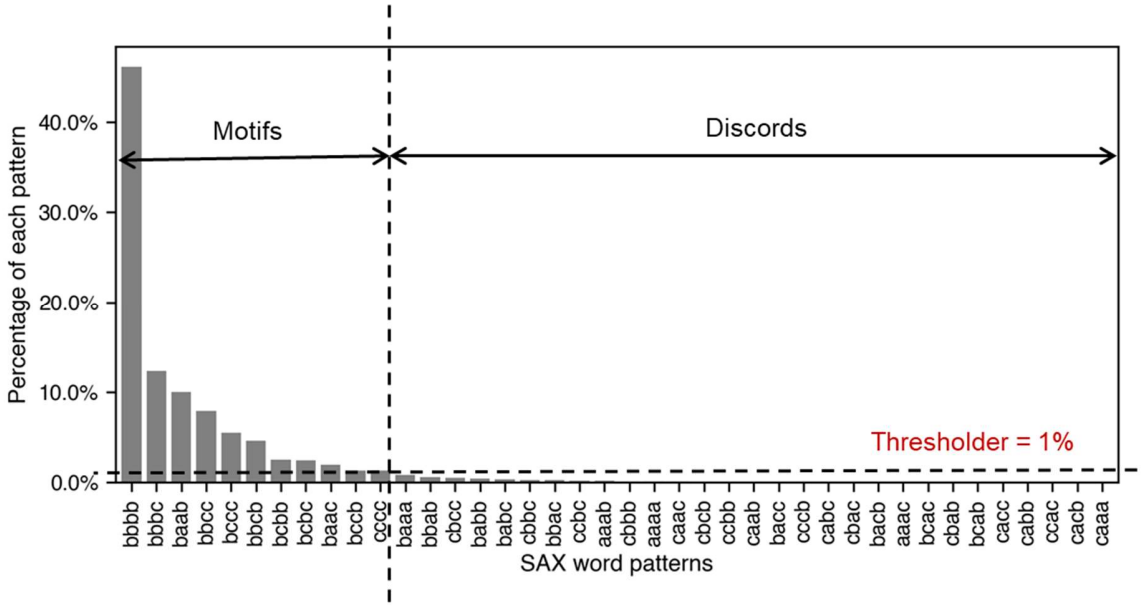

Fig. 4. Distribution of SAX word patterns for 768 households.

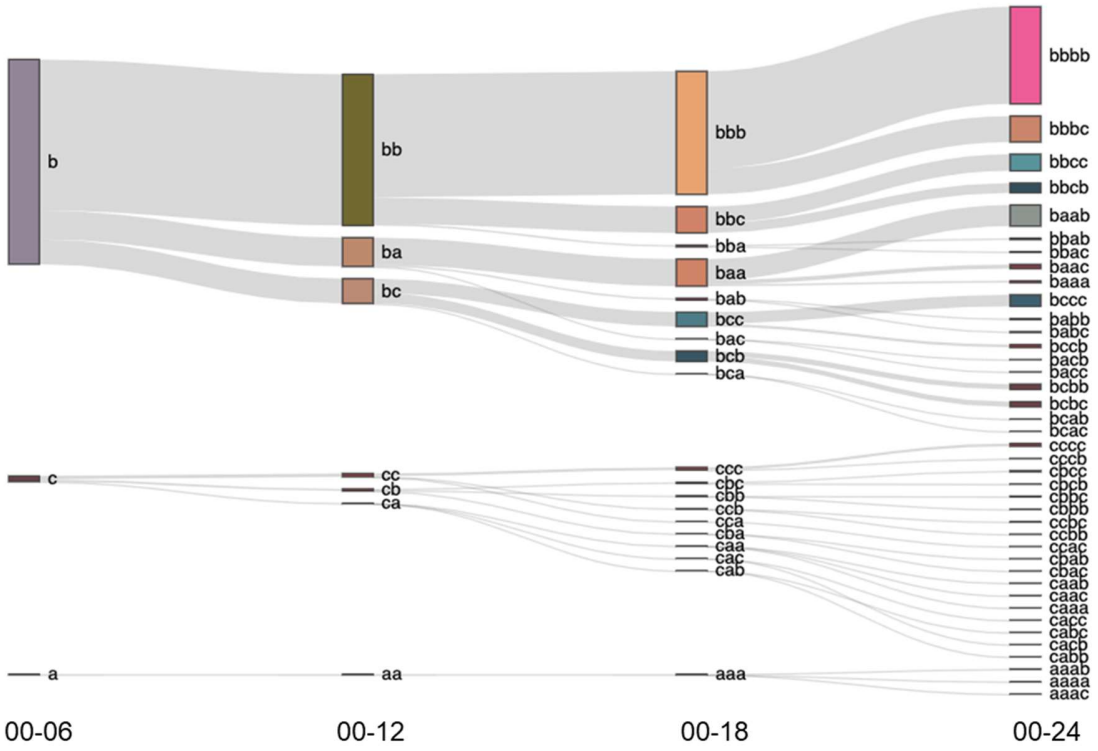

Fig. 5. Augmented suffix tree of daily electricity patterns of 768 households.

4.3. Interpretable feature extraction

With the discovered motif load patterns, a set of interpretable GFs and PFs can be extracted for each pattern. Fig. 6 shows the distributions of the GFs and PFs of year-round daily electricity patterns of all households. Due to the existence of negative net loads for PV households, several statistical features have negative values, including mean value (GF-01), minimum (GF-04), the sum of net loads during business hours (GF-07) and non-business hours (GF-08), and mode of the 5-bin histogram (GF-11). The percentage of values above the mean value (GF-06) is mostly located in the range of [25%, 50%]. As can be seen from the skewness (GF-09) and kurtosis (GF-10) distributions, most distributions of daily electricity consumption profiles are right-tailed and slightly heavy-tailed, relative to a normal distribution. The right-tailed distribution of the daily load profile indicates that the mean value is skewed to the right of the daily median value. From

the temporal feature of GF-12, we can find the longest subsequence where the consecutive value above the mean value is mostly within 4 – 8 hours. Most of the longest periods of successive increase for each daily load pattern are between 2 and 4 hours.

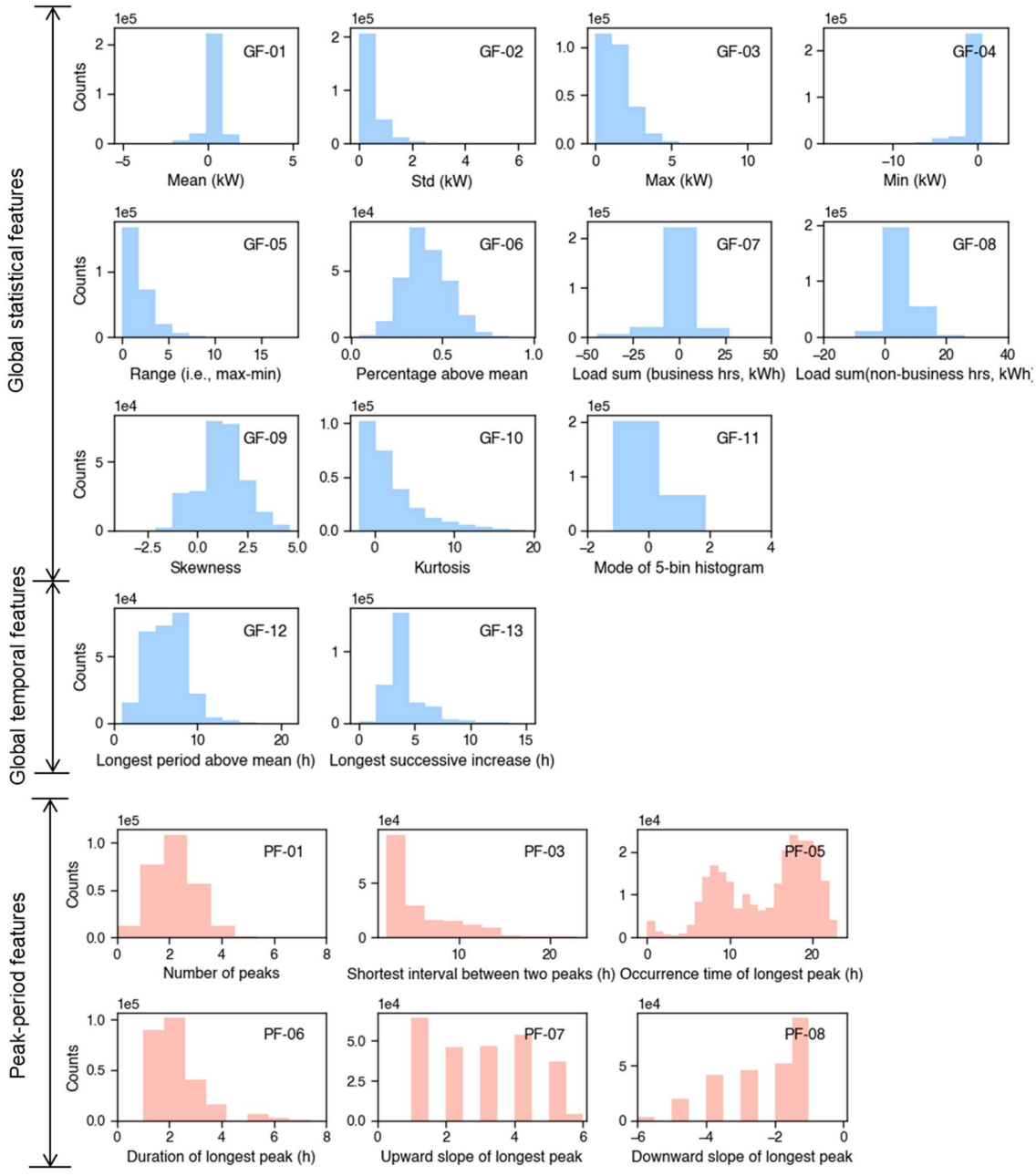

Fig. 6. Distributions of interpretable global and peak-period features for year-round daily electricity patterns of all households. Note that PF-02 and PF-04 are vectors and not feasible for histogram plots.

For peak-period features, since the features of PF-02 (occurrence time of each peak period) and PF-04 (duration of each peak) are vectors and not feasible for histogram plots, the distributions of only 6 peak-period features for daily electricity patterns are plotted in Fig. 6. For most daily load profiles, the number of peaks (PF-01) ranges from 1 to 3. For the profiles with more than one peak, the shortest time interval between peaks (PF-03) is typically less than 5 hours. Regarding the longest peak, they normally occur at around 7:00 AM and 6:00 PM as shown in the distribution of PF-05. Based on the distribution of PF-06, the longest peaks of the majority of the daily load profiles last for 1 – 3 hours. The upward slope and downward slope of the longest peak are majorly in the ranges of [1, 5] and [-4, -1], respectively.

4.4. A two-step feature-based clustering and post-clustering compositional analysis

4.4.1. Monthly feature-based clustering

For both monthly and yearly feature-based clustering, the PCA technique is first applied to reduce the dimensions of the feature matrix from $\mathrm{F} \times \sum_{h=1}^{H} d_h$ to $F_{top} \times \sum_{h=1}^{H} d_h$ for computational efficiency. In the case study, F and H are 21 and 768, respectively. Based on the top feature matrices, the k-means technique is then used for feature-based clustering.

The aim of monthly feature-based clustering is to separate the PV-houses from the mixed households to improve the granularity of clustering results. The daily load profiles in July are used for the monthly feature-based clustering in the case study since PV systems normally generate a considerable amount of solar energy in July, resulting in larger differences in load profiles between PV and non-PV households. Determining the number of principal components (F_{top}) in PCA, and the number of clusters (K) in k-means clustering, is critical to the performance of feature-based clustering results. Fig. 7-a shows the PCA scree plot for 21 interpretable features of daily load

patterns. Based on the cumulative explained variance, PC 1 and PC 2 explain 54.8% of the variance, and PC 1 – PC 10 represent 94.5% of the total variance. The number of the principal components is selected as 10, i.e., $F_{top} = 10$. Table 2 lists the most important feature in each PC for monthly feature-based clustering. It can be seen that both global and peak interpretable features play significant roles in the representation and explanation of the principal components.

After PCA, the k-means clustering is applied based on the top feature matrix. Fig. 7-b shows the trends of Calinski-Harabasz score and Silhouette score when cluster number K ranges from 2 to 10. Based on the elbow point, K is determined to be 5.

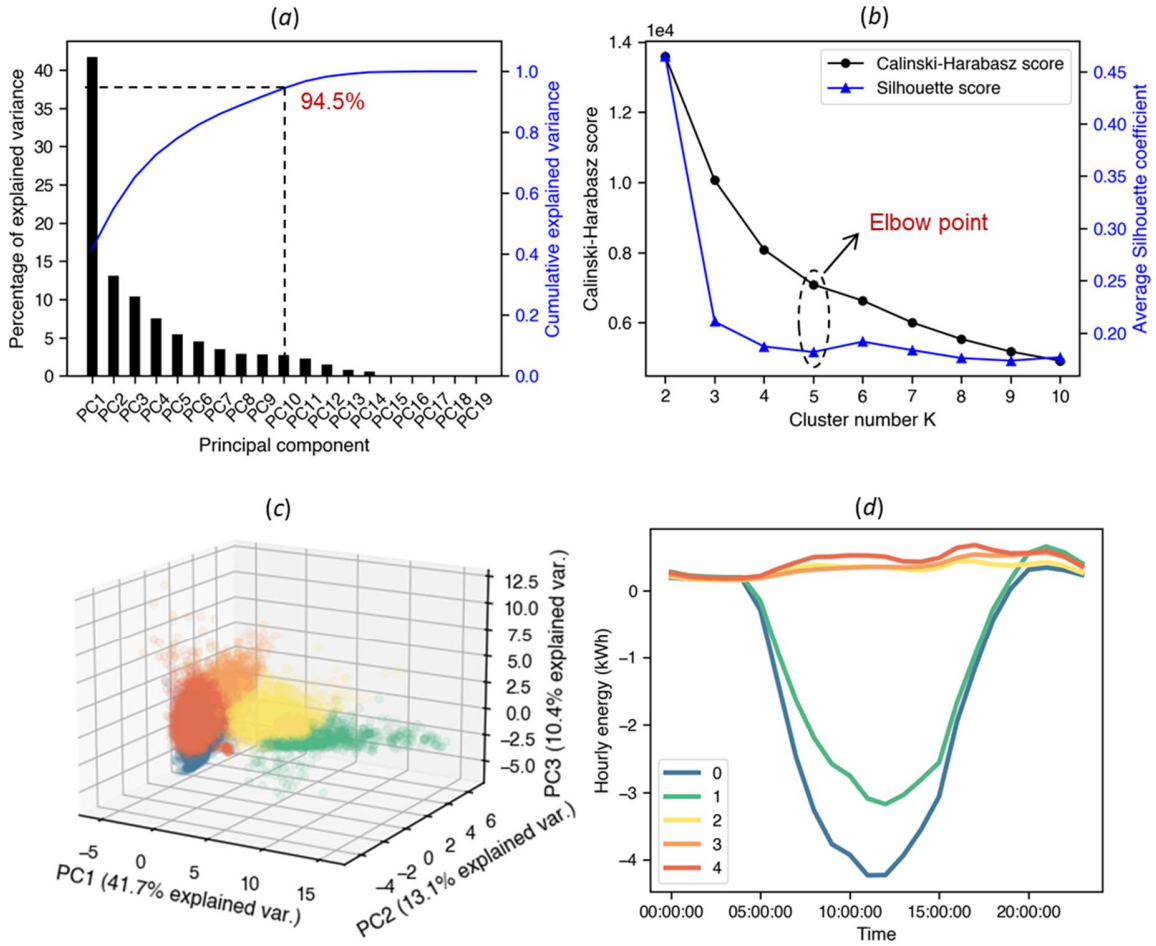

Fig. 7. PCA-based dimensionality reduction and k-means clustering for PV and non-PV households in July: (a) PCA scree plot for 21 interpretable features of daily load patterns; (b) trends of Calinski-Harabasz score and Silhouette score while cluster number *K* ranging from 2 to 10; (c) a scatter plot of three principal components with points being colored based on clustering memberships; (d) centroids of the clusters when $K = 5$.

Table 2. Important features in PCs for monthly feature-based clustering

Principal component	Most important feature in each PC	Percentage of the most important feature
PC1	GF-04	33.89%
PC2	GF-10	45.63%
PC3	GF-03	46.49%
PC4	PF-01	41.67%
PC5	PF-06	42.43%
PC6	PF-03	80.02%
PC7	PF-08	72.09%
PC8	PF-05	51.03%
PC9	GF-11	54.60%
PC10	GF-11	49.65%

Fig. 7-c shows the scatter plot of three principal components (PC 1 vs PC 2 vs PC 3) with points being colored based on clustering memberships. It can be seen that the projected points, representing each daily load profile, can be segmented into 5 areas.. Fig. 7-d shows the centroids of all 5 clusters ($C_{monthly}$), including 2 centroids below y-axis ($C_{monthly}^{negative} = [Cluster\ 0, Cluster\ 1]$) and 3 centroids above y-axis ($C_{monthly}^{positive} = [Cluster\ 2, Cluster\ 3, Cluster\ 4]$). For Household i, the cluster membership of its load patterns, $C_{monthly,i}$, is a subsequence of C_{month} . If $C_{monthly,i}$ contains any element in $C_{monthly}^{negative}$, then Household i is regarded as a PV household; otherwise, it is a non-PV household. For example, for Household 330 in Fig. 8-a, its cluster memberships for the daily profiles in July only contain the centroids in $C_{monthly}^{positive}$, thus Household 330 is a non-PV household. For Household 690 in Fig. 8-b, its cluster

memberships include *Cluster 0*, *Cluster 1*, *Cluster 3*. Since *Cluster 0* and *Cluster 1* belong to $C_{monthly}^{negative}$, Household 690 is a PV household. In the case study, the monthly feature-based clustering can help successfully identify the presence of all of the 154 PV households, and the identification accuracy rate is 100%. The variability and compositional change of the load profiles of a specific household will be analyzed in the yearly feature-based clustering. The monthly feature-based clustering serves only for classifying the PV households and non-PV households.

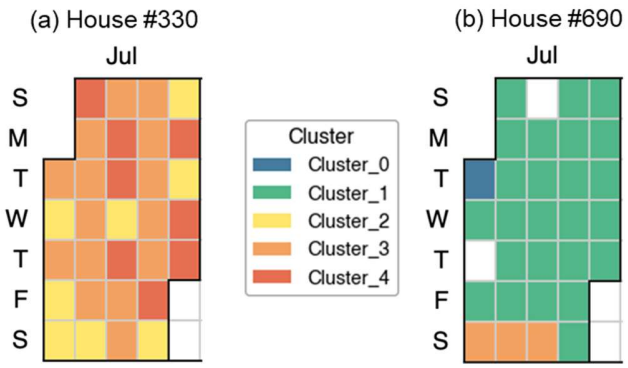

Fig. 8. Examples of the cluster membership distributions of two typical households in July using colored calendar maps: (a) Household 330 (non-PV house); (b) Household 690 (PV house). Note that the white box means the energy profile on that day is not available or filtered.

4.4.2. Yearly feature-based clustering and post-clustering compositional analysis

In this section, the yearly feature-based clustering is performed to get the yearly cluster dictionaries of PV households (C_{yearly}^{PV}) and non-PV households (C_{yearly}^{non-P}). Beyond the clustering analysis, the post-clustering quantitative compositional analysis is further carried out to characterize the variability and compositional change of each household.

- ***Non-PV households***

Fig. 9 shows the results of PCA-based dimensionality reduction and k-means clustering for non-PV households during the whole year. As shown in Fig. 9-a, PC 1 and PC 2 explain 50.6 % of the total variance, and PC 1 – PC 10 represent 94.5% of the total variance. The number of the principal components is selected as 10, i.e., $F_{top} = 10$. Table 3 lists the most important feature in each PC for year-round non-PV household clustering. Both global and peak interpretable features show their effectiveness in the representation and explanation of the major principal components. After PCA, the k-means clustering is applied based on the top feature matrix. Fig. 9-b shows the trends of Calinski-Harabasz score and Silhouette score when cluster number *K* ranges from 2 to 10. Based on the elbow point, *K* is determined to be 6. Fig. 9-c shows the scatter plot of three principal components (PC 1 vs PC 2 vs PC 3) with points being colored based on clustering memberships. The projected points, representing each daily load profile, can be segmented into 6 areas. The centroids of all 6 clusters are also shown in Fig. 9-d. By using the k-means clustering technique, the daily load patterns of all non-PV households can be clearly classified based on the scales and shapes of the patterns. The centroids of Cluster 3 – 5 have similar shapes but largely differ in scales. The centroids of Cluster 0 – 2 are of similar scales but have more differences in shapes.

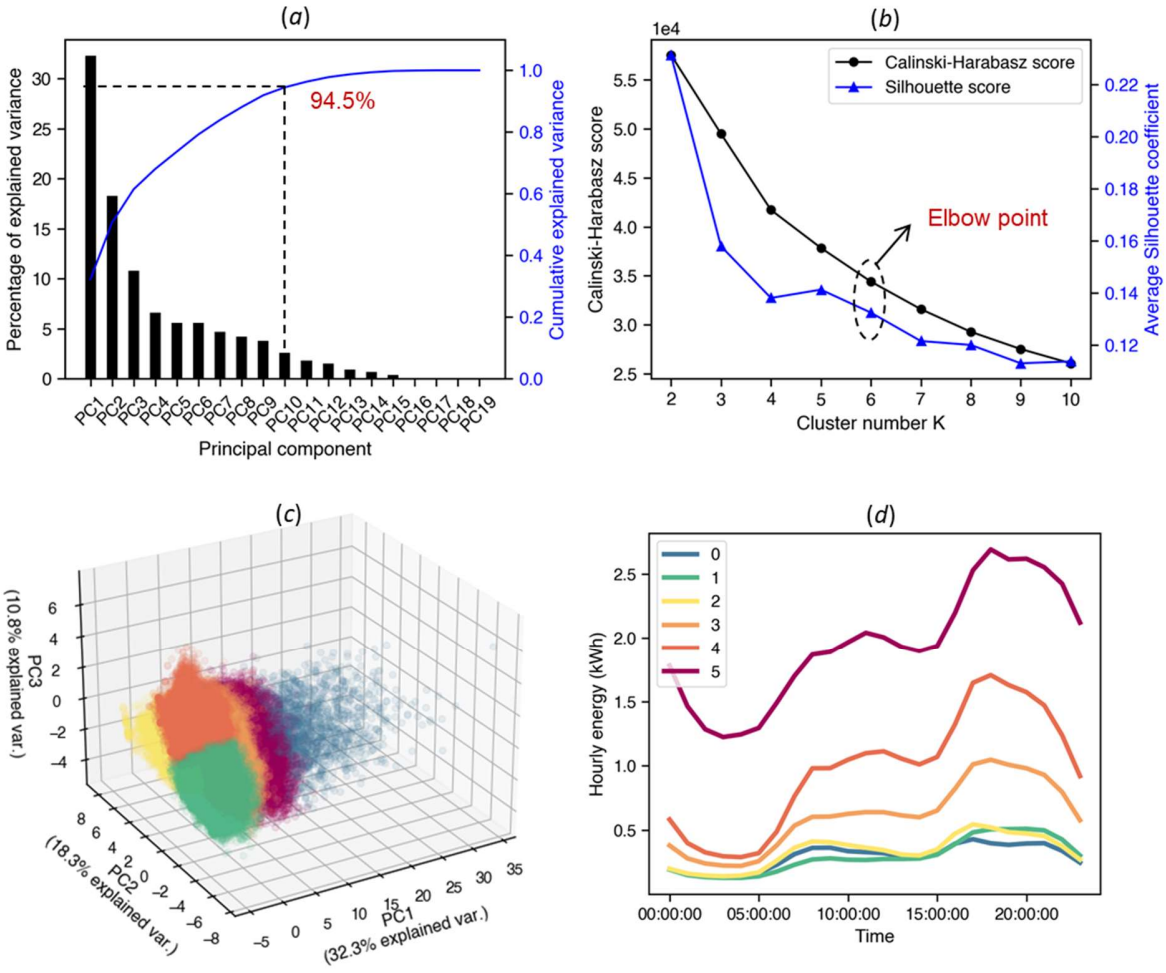

Fig. 9. PCA-based dimensionality reduction and k-means clustering for year-round load patterns of non-PV households: (a) PCA scree plot for 21 interpretable features of daily load patterns; (b) trends of Calinski-Harabasz score and Silhouette score while cluster number *K* ranging from 2 to 10; (c) a scatter plot of three principal components with points being colored based on clustering memberships; (d) centroids of the clusters when $K = 6$.

Table 3. Important features in PCs for year-round non-PV household clustering

Principal component	Most important feature in each PC	Percentage of the most important feature
PC1	GF-01	38.80%
PC2	GF-09	49.88%
PC3	PF-01	49.61%
PC4	PF-06	49.36%
PC5	GF-13	65.13%

PC6	GF-04	66.51%
PC7	PF-03	66.29%
PC8	PF-07	70.17%
PC9	GF-13	49.83%
PC10	PF-05	67.33%

Regarding the energy use variability analysis, the cluster composition for each selected non-PV household is shown in Fig. 10-a and their entropies are plotted in Fig. 10-b. For example, the cluster dictionaries of Household 60 and Household 450 consist of 5 and 3 different clusters, respectively. Household 60, therefore, has a larger value of entropy than Household 450, indicating the energy use patterns of Household 60 are more variable than Household 450. Household 30 and Household 450 have the same number of cluster types. Household 450, however, has a lower variability in energy use than Household 30 because its distribution of cluster memberships is more uneven. The entropy distribution of all 614 non-PV houses is shown in Fig. 10-c.

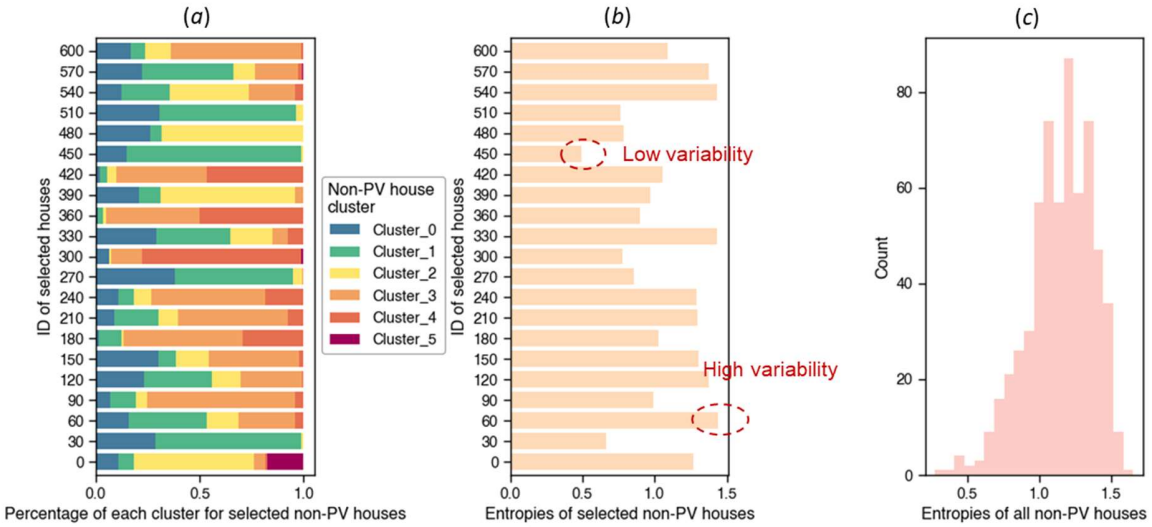

Fig. 10. Variability characterization of non-PV households: (a) cluster composition for each selected non-PV household; (b) entropy of each selected non-PV household; (c) entropies of all non-PV houses.

Besides the variability characterization, the weekly and yearly compositional changes are also quantified using the Aitchison distance. Fig. 11 shows the compositional change analysis for Non-PV Household 60 and Non-PV Household 450. Fig. 11-b shows the cluster composition of Household 60 varies over day of the week. The largest compositional change, i.e., Aitchison distance, exists between Monday and Tuesday and the smallest exists between Saturday and Sunday. The monthly cluster compositions and the trend of yearly compositional change over month of the year are shown in Fig. 11-c. The months in winter have more high-energy profiles than the months in summer. The largest compositional change is the change between October and November. The smallest compositional change exists between December and January.

Compared with Non-PV Household 60 with 5 cluster types, Non-PV Household 450 has only 3 low-energy cluster types (Cluster 0 – 2), indicating a lower variability in energy use. As shown in Fig. 11-e, the cluster composition of Household 450 varies over day of the week. The largest compositional change, i.e., Aitchison distance, exists between Tuesday and Wednesday and the smallest exists between Sunday and Monday. The monthly cluster compositions of Household 450 and the trend of yearly compositional change over month of the year are shown in Fig. 11-f. For Household 450, the months in summer have more high-energy profiles than the month in winter. The largest compositional change is the change between March and April. Since the cluster compositions are the same from October to December, the change between October and November and the change between November and December are the smallest compositional changes, i.e., 0.

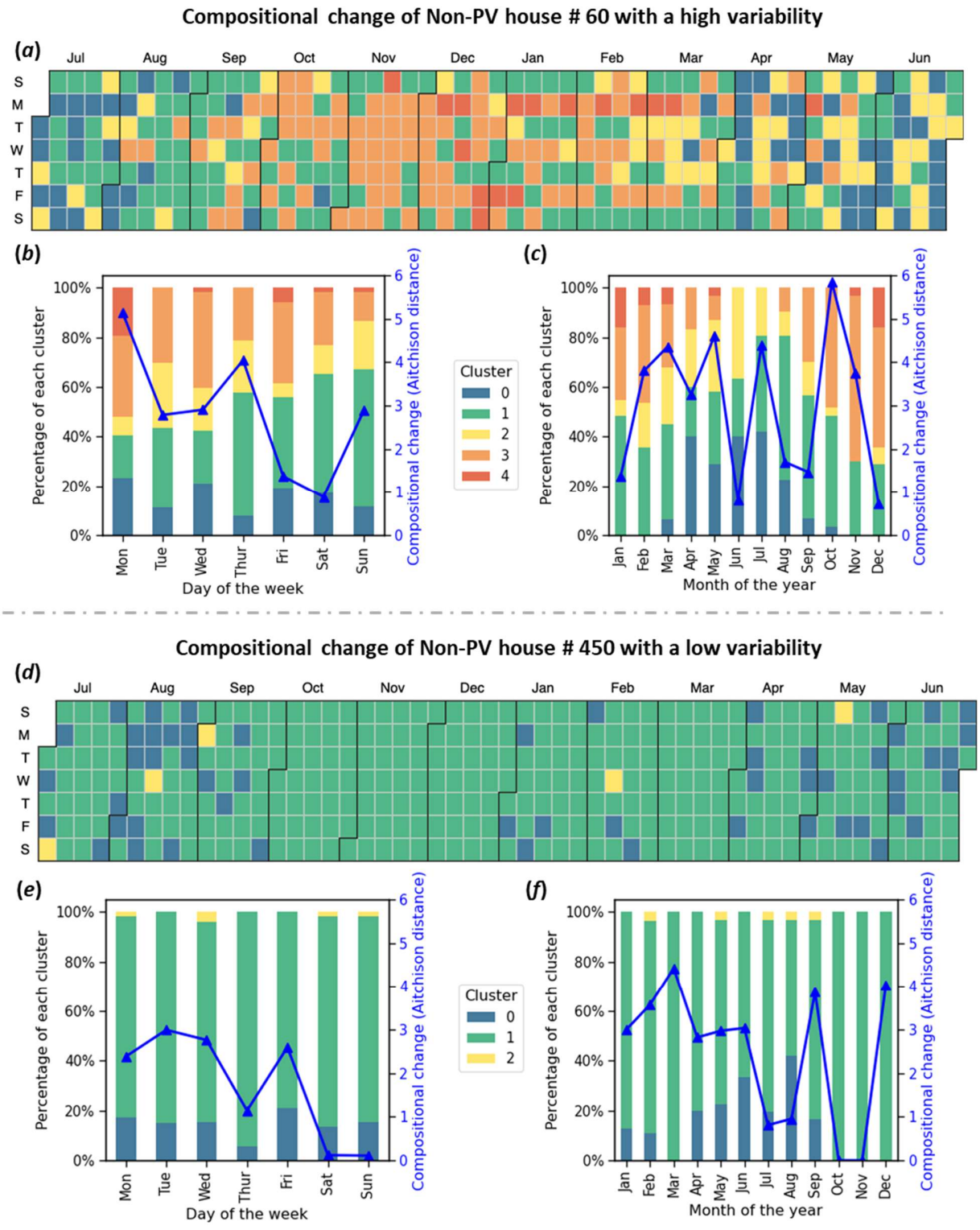
Compositional change of Non-PV house # 60 with a high variability
(a)
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
S
M
T
W
T
F
S
(b)
Percentage of each cluster
Compositional change (Aitchison distance)
Cluster
0
1
2
3
4
Mon
Tue
Wed
Thur
Fri
Sat
Sun
Day of the week
(c)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month of the year
Compositional change of Non-PV house # 450 with a low variability
(d)
(e)
(f)

Fig. 11. Compositional changes of the Non-PV Household # 60 and # 450: (a and d) year-round distribution of daily cluster membership; (b and e) weekly compositional change over day of the week (Note that the last point represents the compositional change between Sunday and Monday); (c and f) yearly compositional change over month of the year.

- ***PV households***

Fig. 12 shows the results of PCA-based dimensionality reduction and k-means clustering for year-round load patterns of PV households. As shown in Fig. 12-a, PC 1 and PC 2 explain 56.7 % of the total variance, and PC 1 – PC 10 represent 95.96% of the total variance. The number of the principal components is selected as 11, i.e., $F_{top} = 11$. Table 4 lists the most important feature in each PC for year-round PV household clustering. Both global and peak interpretable features show their effectiveness in the representation and explanation of the major principal components. After PCA, the k-means clustering is applied based on the top feature matrix. Fig. 12-b shows the trends of Calinski-Harabasz score and Silhouette score when cluster number *K* ranges from 2 to 10. Based on the elbow point, *K* is determined to be 6. Fig. 12-c shows the scatter plot of three principal components (PC 1 vs PC 2 vs PC 3) with points being colored based on clustering memberships. The projected points, representing each daily load profile, can be segmented into 6 areas. The centroids of all 6 clusters are also shown in Fig. 12-d. It can be seen that by using the k-means clustering technique, the daily load patterns of all PV households can be clearly segmented based on the scales and shapes of the patterns. The centroids of Cluster 0 – 5 have similar shapes but largely differ in scales. Note that the clustering results, including numbers and shapes of cluster centroids, are influenced by the ratings of PV systems on a particular distribution network.

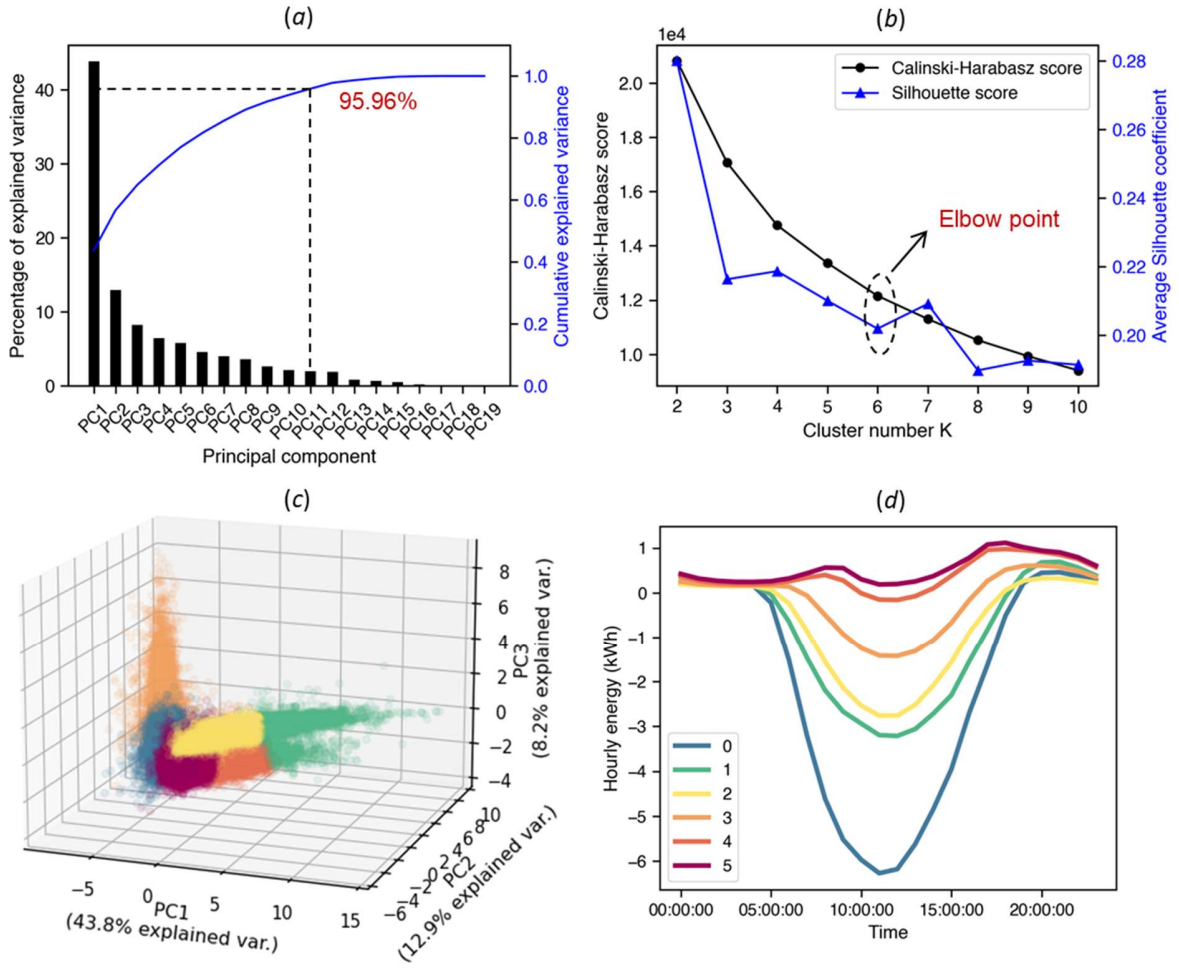

Fig. 12. PCA-based dimensionality reduction and k-means clustering for year-round load patterns of PV households: (a) PCA scree plot for 21 interpretable features of daily load patterns; (b) trends of Calinski-Harabasz score and Silhouette score while cluster number *K* ranging from 2 to 10; (c) a scatter plot of three principal components with points being colored based on clustering memberships; (d) centroids of the clusters when $K = 6$.

Table 4. Important features in PCs for year-round PV household clustering

Principal component	Most important feature in each PC	Percentage of the most important feature
PC1	GF-01	32.15%
PC2	GF-12	37.91%
PC3	PF-05	53.58%
PC4	GF-03	46.45%
PC5	PF-03	59.52%
PC6	GF-11	46.11%

PC7	GF-11	62.56%
PC8	GF-13	75.53%
PC9	GF-10	53.26%
PC10	PF-07	68.87%
PC11	GF-12	53.15%

Regarding the energy use variability analysis, the cluster composition for each selected PV household is shown in Fig. 13-a and their entropies are plotted in Fig. 13-b. For example, the cluster dictionaries of both PV Household 0 and PV Household 100 consist of 6 different clusters, but the cluster membership distribution of PV Household 0 is more uniform. PV Household 0, therefore, has a larger value of entropy compared with PV Household 100, which indicates the energy use patterns of PV Household 0 are more variable than PV Household 100. The entropy distribution of all 154 PV houses is shown in Fig. 13-c.

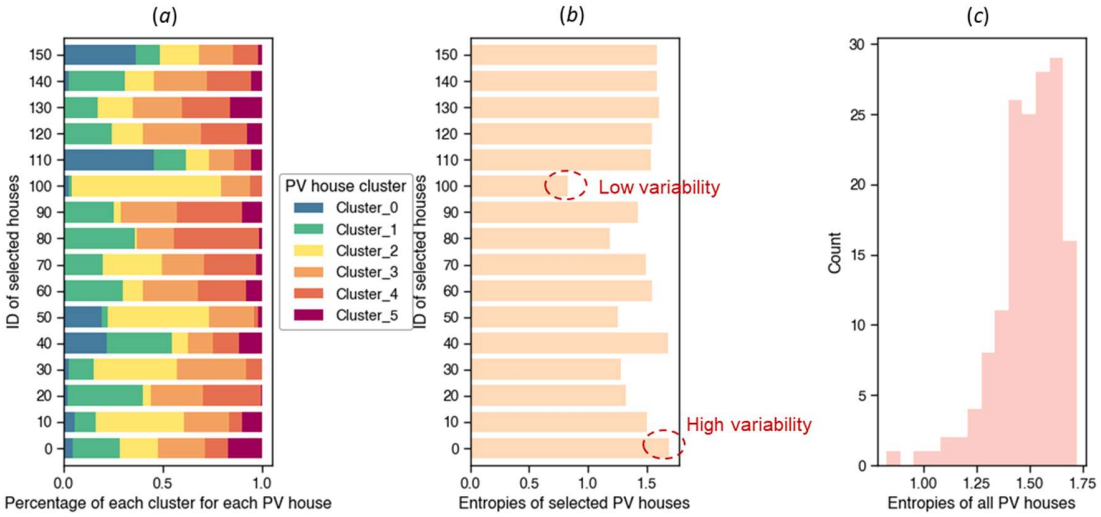

Fig. 13. Variability characterization of PV households: (a) cluster composition for each selected PV household; (b) entropy of each selected PV household; (c) entropies of all PV households.

Besides the variability characterization, the weekly and yearly compositional changes are also quantified using the Aitchison distance. Fig. 14 shows the compositional change analysis for PV

Household 0 and PV Household 100. It can be seen from Fig. 14-b that the cluster composition of PV Household 0 varies over day of the week. The largest compositional change, i.e., Aitchison distance, exists between Friday and Saturday, and the smallest exists between Tuesday and Wednesday. The monthly cluster compositions and the trend of yearly compositional change over month of the year are shown in Fig. 14-c. For PV Household 0, the months in winter have more high-energy profiles than the months in summer and the largest compositional change is the change between February and March. The smallest compositional change exists between April and May, which means the energy use patterns are similar in the two months.

Compared with PV Household 0, PV Household 100 has the same number of cluster types, but its distribution of cluster memberships is less uniform, indicating a lower variability in energy use. As can be seen from Fig. 14-e, the cluster composition of PV Household 100 varies over day of the week. The largest compositional change, i.e., Aitchison distance, exists between Sunday and Monday and the smallest exists between Saturday and Sunday. The monthly cluster compositions of PV Household 100 and the trend of yearly compositional change over month of the year are shown in Fig. 14-f. For PV Household 100, the months in winter have more high-energy profiles than the month in summer. The largest compositional change is the change between April and May. The smallest compositional change exists between August and September, which means the energy use patterns are similar in the two months.

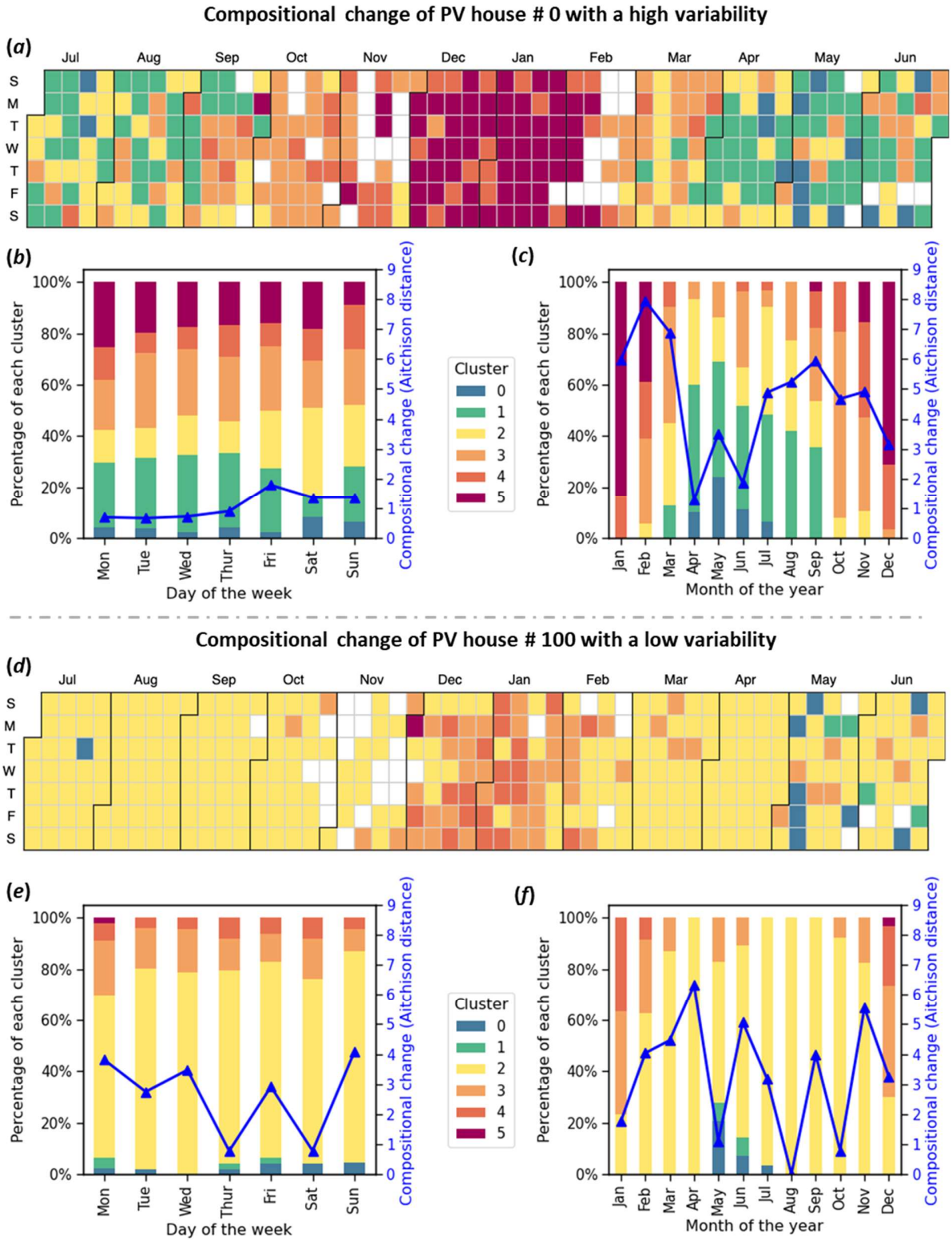
Compositional change of PV house # 0 with a high variability
(a)
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
S
M
T
W
T
F
S
(b)
Percentage of each cluster
Compositional change (Aitchison distance)
Day of the week
Mon
Tue
Wed
Thur
Fri
Sat
Sun
Cluster
0
1
2
3
4
5
(c)
Month of the year
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Compositional change of PV house # 100 with a low variability
(d)
(e)
(f)

Fig. 14. Compositional change of the PV Household # 0 and # 100: (a and d) year-round distribution of daily cluster membership (Note that the white box means the energy profile on that day is not available or filtered); (b and e) weekly compositional change over day of the week (Note that the last point represents the compositional change between Sunday and Monday); (c and f) yearly compositional change over month of the year.

5. Conclusions

Energy use on the demand side is changing and in order to accommodate it within the wider power system context, it needs to be better understood in terms of its magnitude, heterogeneity, and diversity. Embedded unauthorized PVs pose hidden threats to distribution networks in terms of voltage regulation, frequency control, circuit reconfiguration, and back-feeding flow. To automate the identification of this on a large scale, this paper developed a holistic data analytics approach to classifying and characterizing the intra-day load curves of PV and Non-PV households. The developed interpretable global and peak-period features enable the extraction of the physical and interpretable information in the daily load patterns. The adopted symbolic representation technique can help quickly capture the dynamics of load patterns during peak periods which are important from an energy flexibility perspective. By performing PCA-based dimensionality reduction, the top discriminating features can be identified and readily used for clustering purposes. A two-step feature-based clustering can help efficiently classify and characterize the electricity load profiles of PV and non-PV households. The post-clustering compositional analysis can quantitatively characterize the energy consumption variability and compositional change over a week/year for each household.

The proposed holistic data analytics method in this study enables electricity suppliers to fulfill two tasks: *PV household identification* and *feature-based load pattern categorization*. The effective classification of PV and non-PV households allows them to identify the penetration rate of PV

systems in a specific area under a secondary substation, which can assist them with voltage control and ancillary services. Besides, the load pattern categorization can help better understand the demand characteristics of different cohorts of customers, network utilization, and changes in total load growth on a distribution network. The enhanced insights can help electricity suppliers improve demand-side management performance, load forecasting accuracy, and bespoke tariff formulation on low-voltage distribution networks. In the future, we will develop energy consumption models for each category of customers to improve the load forecasting accuracy and further apply the models to model-based demand-side management.

Code Availability

The code for the process of interpretable feature extraction has been packaged and distributed on both GitHub (https://github.com/chacehoo/IFEEL) and PyPI (https://pypi.org/project/ifeel/).

Acknowledgments

The authors would like to acknowledge AECOM Building Engineering, Department of Energy and Climate Change, Centre for Sustainable Energy, and the UK Data Archive for providing the data used in this paper. This work was financially supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under the grant (EP/S030131/1).

References

[1] Siano P. Demand response and smart grids—A survey. Renewable and Sustainable Energy Reviews. 2014;30:461-78.
[2] Neves D, Brito MC, Silva CA. Impact of solar and wind forecast uncertainties on demand response of isolated microgrids. Renewable Energy. 2016;87:1003-15.
[3] Khan MW, Wang J, Ma M, Xiong L, Li P, Wu F. Optimal energy management and control aspects of distributed microgrid using multi-agent systems. Sustainable Cities and Society. 2019;44:855-70.
[4] Gelazanskas L, Gamage KAA. Demand side management in smart grid: A review and proposals for future direction. Sustainable Cities and Society. 2014;11:22-30.

[5] Aduda KO, Labeodan T, Zeiler W, Boxem G, Zhao Y. Demand side flexibility: Potentials and building performance implications. Sustainable Cities and Society. 2016;22:146-63.
[6] Hu M, Xiao F, Jørgensen JB, Wang S. Frequency control of air conditioners in response to real-time dynamic electricity prices in smart grids. Applied Energy. 2019;242:92-106.
[7] Hu M, Xiao F, Wang S. Neighborhood-level coordination and negotiation techniques for managing demand-side flexibility in residential microgrids. Renewable and Sustainable Energy Reviews. 2021;135.
[8] Quilumba FL, Lee W-J, Huang H, Wang DY, Szabados RL. Using Smart Meter Data to Improve the Accuracy of Intraday Load Forecasting Considering Customer Behavior Similarities. IEEE Transactions on Smart Grid. 2015;6:911-8.
[9] Kwac J, Flora J, Rajagopal R. Lifestyle Segmentation Based on Energy Consumption Data. IEEE Transactions on Smart Grid. 2018;9:2409-18.
[10] Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy. 2012;42:68-80.
[11] Axon CJ, Darby SJ, Granell R, Hobson PR, Layberry RL, Pisica I, et al. Towards an understanding of dynamic energy pricing and tariffs. 2012 47th International Universities Power Engineering Conference (UPEC)2012. p. 1-5.
[12] BEIS. Smart Meter Statistics in Great Britain: Quarterly Report to end September 2020. Department for Business, Energy, and Industrial Strategy. 2020.
[13] Wang Y, Chen Q, Hong T, Kang C. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. IEEE Transactions on Smart Grid. 2019;10:3125-48.
[14] Zhang X, Grijalva S. A Data-Driven Approach for Detection and Estimation of Residential PV Installations. IEEE Transactions on Smart Grid. 2016;7:2477-85.
[15] Razavi S-E, Rahimi E, Javadi MS, Nezhad AE, Lotfi M, Shafie-khah M, et al. Impact of distributed generation on protection and voltage regulation of distribution systems: A review. Renewable and Sustainable Energy Reviews. 2019;105:157-67.
[16] Malof JM, Bradbury K, Collins LM, Newell RG. Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. Applied Energy. 2016;183:229-40.
[17] Wang F, Li K, Wang X, Jiang L, Ren J, Mi Z, et al. A Distributed PV System Capacity Estimation Approach Based on Support Vector Machine with Customer Net Load Curve Features. Energies. 2018;11.
[18] Zhou K-l, Yang S-l, Shen C. A review of electric load classification in smart grid environment. Renewable and Sustainable Energy Reviews. 2013;24:103-10.
[19] Yi W, Qixin C, Chongqing K, Mingming Z, Ke W, Yun Z. Load profiling and its application to demand response: A review. Tsinghua Science and Technology. 2015;20:117-29.
[20] do Carmo CMR, Christensen TH. Cluster analysis of residential heat load profiles and the role of technical and household characteristics. Energy and Buildings. 2016;125:171-80.
[21] Gianniou P, Liu X, Heller A, Nielsen PS, Rode C. Clustering-based analysis for residential district heating data. Energy Conversion and Management. 2018;165:840-50.
[22] Zhan S, Liu Z, Chong A, Yan D. Building categorization revisited: A clustering-based approach to using smart meter data for building energy benchmarking. Applied Energy. 2020;269.
[23] Chaouch M. Clustering-Based Improvement of Nonparametric Functional Time Series Forecasting: Application to Intra-Day Household-Level Load Curves. IEEE Transactions on Smart Grid. 2014;5:411-9.
[24] Jota PRS, Silva VRB, Jota FG. Building load management using cluster and statistical analyses. International Journal of Electrical Power & Energy Systems. 2011;33:1498-505.

[25] Verdu SV, Garcia MO, Franco FJG, Encinas N, Marin AG, Molina A, et al. Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters. IEEE PES Power Systems Conference and Exposition, 20042004. p. 899-906 vol.2.
[26] Räsänen T, Voukantsis D, Niska H, Karatzas K, Kolehmainen M. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Applied Energy. 2010;87:3538-45.
[27] Granell R, Axon CJ, Wallom DCH. Clustering disaggregated load profiles using a Dirichlet process mixture model. Energy Conversion and Management. 2015;92:507-16.
[28] Chicco G, Napoli R, Piglione F. Comparisons Among Clustering Techniques for Electricity Customer Classification. IEEE Transactions on Power Systems. 2006;21:933-40.
[29] Granell R, Axon CJ, Wallom DCH. Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles. IEEE Transactions on Power Systems. 2015;30:3217-24.
[30] Westermann P, Deb C, Schlueter A, Evins R. Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data. Applied Energy. 2020;264.
[31] Granell R, Axon CJ, Wallom DCH. Predicting winning and losing businesses when changing electricity tariffs. Applied Energy. 2014;133:298-307.
[32] Timmer J, Gantert C, Deuschl G, Honerkamp J. Characteristics of hand tremor time series. Biological Cybernetics. 1993;70:75-80.
[33] Patel P, Keogh E, Lin J, Lonardi S. Mining motifs in massive time series databases. 2002 IEEE International Conference on Data Mining, 2002 Proceedings2002. p. 370-7.
[34] Lin J, Keogh E, Wei L, Lonardi S. Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery. 2007;15:107-44.
[35] Keogh E, Lin J, Fu A. HOT SAX: efficiently finding the most unusual time series subsequence. Fifth IEEE International Conference on Data Mining (ICDM'05)2005. p. 8 pp.
[36] Miller C, Nagy Z, Schlueter A. Automated daily pattern filtering of measured building performance data. Automation in Construction. 2015;49:1-17.
[37] Hu M, Ge D, Wallom D. A Python package for Interpretable Feature Extraction of Electricity Loads (IFEEL). 2020. Available: https://pypi.org/project/ifeel/
[38] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374:20150202.
[39] Sophian A, Tian GY, Taylor D, Rudlin J. A feature extraction technique based on principal component analysis for pulsed Eddy current NDT. NDT & e International. 2003;36:37-41.
[40] Lakhina S, Joseph S, Verma B. Feature reduction using principal component analysis for effective anomaly–based intrusion detection on NSL-KDD. 2010.
[41] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53-65.
[42] Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics. 1974;3:1-27.
[43] Aitchison J. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological). 1982;44:139-60.
[44] Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V. Logratio Analysis and Compositional Distance. Mathematical Geology. 2000;32:271-5.
[45] Buccianti A, Lima A, Albanese S, De Vivo B. Measuring the change under compositional data analysis (CoDA): Insight on the dynamics of geochemical systems. Journal of Geochemical Exploration. 2018;189:100-8.

[46] Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. Mathematical Geology. 2003;35:253-78.
[47] AECOM Building Engineering. Energy Demand Research Project: Early Smart Meter Trials, 2007-2010. Colchester, Essex: UK Data Archive; 2014.
[48] Sandia National Lab. PV Toolbox. 2020. Available: https://pvpmc.sandia.gov/applications/pv_lib-toolbox/
[49] Typical Meteorological Year (TMY). 2020. Available: https://e3p.jrc.ec.europa.eu/articles/typical-meteorological-year-tmy