# New Methods for Clustering District Heating Users Based on Consumption Patterns

Chao Wang, Yuyan Du, Hailong Li, Fredrik Wallin, and Geyong Min

**Abstract**

Understanding energy users' consumption patterns benefits both utility companies and consumers as it can support improving energy management and usage strategies. The rapid deployment of smart metering facilities has enabled the analysis of consumption patterns based on high-precision real usage data. This paper investigates data-driven unsupervised learning techniques to partition district heating users into separate clusters such that users in the same cluster possess similar consumption pattern. Taking into account the characteristics of heat usage, three new approaches of extracting pattern features from consumption data are proposed. Clustering algorithms with these features are executed on a real-world district heating consumption dataset. The results can reveal typical daily consumption patterns when the consumption linearly related to ambient temperature is removed. Users with heat usages that are highly imbalanced within a certain period of time or are highly consistent with the utility heat production load can also be grouped together. Our methods can facilitate gaining better knowledge regarding the behaviors of district heating users and hence can potentially be used to formulate new pricing and energy reduction solutions.

**Index Terms**

District heating, user clustering, energy consumption pattern, feature extraction

## I. Introduction

Population growth and economic development have led to a surge in the global demand for energy in recent years. The consequent energy crisis and environment pollution have also widely caught public attentions. The building sector is regarded as one of the largest emitters of $CO_2$ to the global atmosphere [1]. The International Energy Agency (IEA) estimates that, worldwide, buildings represent 32% of total final energy consumption [2], 34.8% of which is used for space heating, ventilation and airconditioning (HVAC) [3]. Therefore, saving energy in buildings is crucial to achieve significant reduction of greenhouse gas emission.

District heating (DH) is characterized by high energy efficiency and low environmental pollution among different methods for heat supply. Understanding the behaviors of consumers and patterns of heat consumption is of importance to ensure the reliability and efficiency of heat management [4]. Such information can be used by utility companies to predict energy demands, identify abnormal activities, suspect energy fraud, and improve demand-side management and tariff settings. It can also empower consumers to optimize their behaviors and renovate their facilities in order to reduce energy cost [5]. Along with the development of automated smart metering technologies, a large amount of high-precision heat consumption records have become available, which opens a new era for extracting valuable information from real-world data [6].

Clustering is one of the most commonly adopted unsupervised machine learning methods that exploit knowledge from data.. Through clustering, energy users are partitioned into different groups. In the same

C. Wang and Y. Du are with the Department of Information and Communication Engineering, Tongji University, Shanghai, China. H. Li and F. Wallin are with the Future Energy, Mälardalens Högskola, Västerås, Sweden. G. Min is with the Department of Computer Science, University of Exeter, Exeter, UK. C. Wang is also with the Department of Computer Science, University of Exeter, Exeter, UK.

H. Li is the correspondence author. Email: hailong.li@mdh.se.

This paper is a substantial extension of a short version paper presented at ICAE'2018, Hong Kong, China, 22-25 Aug. 2018.

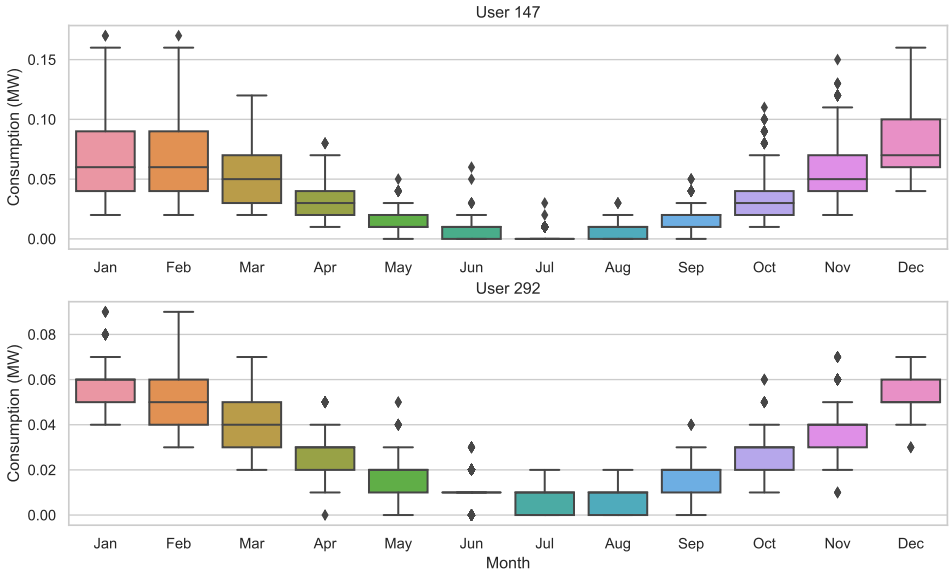
Abbreviations: BIC: Bayes information criterion; CPC: consumption-production consistency; CV: coefficient of variation; DDC: discretized duration curve; DH: district heating; DLP: daily load profile; GMM: Gaussian mixture model; LOF: local outlier factor; MDLP: modified daily load profile; NMI: normalized mutual information; PCC: Pearson correlation coefficient; PMF: probability mass function

group, users possess a certain similar pattern, which is different from that of users in other groups. Clustering *electricity* usage patterns has attracted a huge amount of recent research interests. The most commonly considered pattern feature for clustering is the daily load profile (DLP), i.e., the typical electricity usage variation tendency in a 24-hour period. For instance, [7] generates the DLP and conducts a comparison of the clustering results attained using the hierarchical, k-means, and fuzzy c-means algorithms. A disaggregation analysis on smart metering data is conducted by applying the fuzzy c-means clustering algorithm in [8]. [9] focuses on the DLP of family houses and identifies several human-related factors that affect the shape of the average electricity usage profile. Some works also study variation patterns in longer periods, e.g., the weekly load profile [10] and monthly load profile [11]. Due to the fact that a user may have very disparate weekday and weekend consumption patterns, [5] considers excluding weekends before establishing DLP. The impact of temporal resolution and normalization methods on clustering results is investigated in [12] [13]. Different clustering algorithms can be applied to partition users, including, e.g., $k$-means, fuzzy c-means, $k$-medoids, self organizing maps [14], and spectral clustering [15], to name a few. Combining consumption data with commercial/cartographic data [15] and even user surveys [16] to facilitate clustering is also feasible. Clustered electricity load profiles are shown to have a wide range of applications, including finding the right segments in pricing schemes [17], demand-side management [18], and improving forecasting [19], etc.

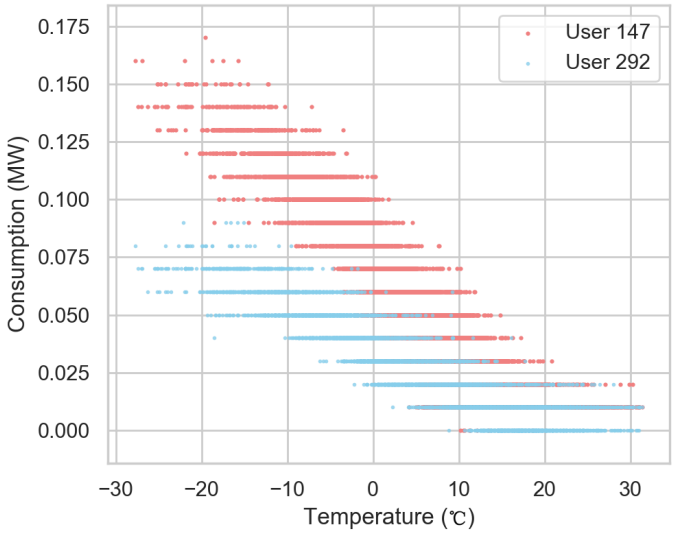
Heat consumption patterns may exhibit very different characteristics from that of electricity, which makes the aforementioned clustering methods hard to apply directly. Heat usage is in general primarily dedicated to space heating. The relatively simpler function compared to electricity consumption causes the usage pattern to be insensitive to other factors such as human activities. This enables demand forecasting to be conducted with relatively high accuracy, through considering only outdoor weather, consumers' desire of comfortable indoor temperature, and physical factors including building function, structure and layout. For example, [20] conducts heat demand prediction of buildings in DH systems using the Gaussian mixture model (GMM) with temperature and building function information. A model based on Elman neural networks is adopted in [21] to explore the key factors in heat demand forecasting. [22] develops a two-step prediction approach by taking engineering estimates as prior knowledge. [23] introduces a procedure for developing energy consumption quota of public buildings. [24] applies statistical models including linear regression, random forest, and support vector regression, to predict the energy usage of buildings in New York with a number of physical and spatial attributes. Most of these works take advantages of the fact that space heating accounts for a large amount of heat usage. However, this also brings challenges for clustering analysis, as it leads to difficulties in distinguishing different users' individual usage patterns.

Fig. 1 displays the heat consumption (generated from a total of 8328 hourly records collected in 347 days in 2010) of two DH users in Sweden. From the boxplot of the daily consumption across different months in Fig. 1(a), it can be clearly seen that the heat usage in winter season (e.g., January, February, November, and December) was significantly higher than that in summer season (e.g., June, July and August). Hence, directly use the average of daily consumption data to form the DLP and conduct clustering would cause the clustering result to be biased towards cold days' pattern. Fig. 1(b) displays the relationship between consumption and outdoor temperature. A strong linear correlation is observable: The Pearson correlation coefficients (PCCs) for the two example users, User 147 and User 292, are $-0.832$ and $-0.811$, respectively. As all users in the same region experience almost the same weather condition, their heat usage would change following a similar tendency, which makes it hard to distinguish actual user-specific patterns.

In addition, the DLP exhibits only a usage variation pattern on a daily basis. It may not be able to fully reflect the factors that utility companies, i.e., DH companies, concern in their business activities such as system operation, production management, pricing policy design, and budget planning. For instance, usage characteristics related to, e.g., the ratio of peak to average consumptions, the accumulated duration that each consumption level occurs, and the network peak demand, within a relatively long period, cannot be revealed by the DLP. Furthermore, the averaging operation conducted in the process of generating DLP blurs the distinctions between each individual day's usage. It is not helpful when one intends to measure




(a)




(b)

Fig. 1: Heat consumption data of two example DH users. (a) Daily consumptions in each month. (b) Hourly consumption versus outdoor temperature. User147 has Pearson correlation coefficient (PCC) of $-0.832$ and User 292 has PCC of $-0.811$.

how much pressure a DH user's heat demand puts on the DH company's production capacity. For example, if a user always has high heat consumption when the DH company has to raise its generation output to satisfy a high demand from its customers, the user would have a high contribution to the network peak load. This may potentially result in a large amount of extra operation expenses for the company. The DLP is not capable of reflecting such differences between DH users. Therefore, new pattern features for performing DH user clustering should be considered.

DH user clustering has rarely been investigated in the literature. DH companies usually categorize their customers into different groups according to the main function of users, such as multi-family residence, office building, industry factory, and commercial building, etc., based on which different price models can be adopted [25]. However, user function cannot reflect heat consumption pattern accurately. To fill in the knowledge gap, this paper provides a study of an exploratory sort on DH user clustering. Our

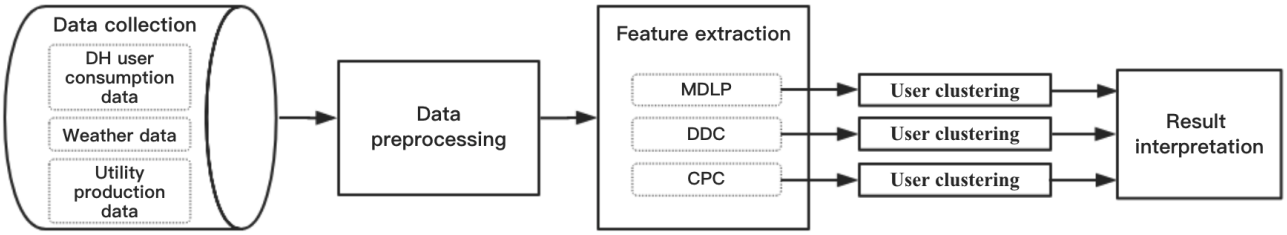



Fig. 2: Flowchart for clustering district heating users.

contributions are twofold:

*1) New methods of extracting representative pattern features from heat usage record data:* Comparing with electricity usage, heat consumption has a stronger relationship with ambient outdoor temperature. Such a relationship may conceal the consumption patterns dependent on consumers' behaviors. In order to understand more deeply the user-specific daily usage variation tendencies, a method is proposed to separate the usage that is linearly affected by ambient temperature from the total consumption. Moreover, long-term user behavior and its impact on network load demand are also of interest. Therefore, another two methods are proposed in order to extract the features that can respectively reflect whether a DH user has balanced usage of different consumption levels in a long period of time, and whether its consumption is always consistent with the utility heat production load.

*2) New methods of clustering DH users:* Existing methods and results regarding DH user clustering are limited. Using the above pattern features, one can cluster DH consumers and understand their heat usage patterns from new perspectives. Through a case study on a practical dataset, it is shown that these new approaches can provide richer information than the traditional way of using function to categorize DH users. The results can potentially be used to support developing smart energy management and usage solutions. For example, grouping users with the same consumption pattern can facilitate precise load demand prediction. Understanding the typical usage variation tendency of customers can also facilitate time-of-day pricing design. Users who are identified to have unbalanced usage or high contribution to network peak load can be encouraged to improve their consuming behaviors.

The remainder of the paper is organized as follows. Section II briefly explains our methodology and work flow. The detailed approaches of extracting the proposed pattern features from heat consumption data are elaborated in Section III. In Section IV, the results of DH user clustering using a practical dataset are presented. Finally, our paper is concluded in Section V.

*Notation*: mean$\{\boldsymbol{v}\}$ and std$\{\boldsymbol{v}\}$ represent mean and standard deviation of data vector $\boldsymbol{v}$, respectively. min$\{\boldsymbol{v}\}$ and max$\{\boldsymbol{v}\}$ respectively find the minimum and maximum element in $\boldsymbol{v}$. cov$\{\boldsymbol{v}, \boldsymbol{w}\}$ represents the covariance between $\boldsymbol{v}$ and another data vector $\boldsymbol{w}$. $\lfloor \cdot \rfloor$ denotes the floor operation. log represents base-$e$ logarithm operation. $\boldsymbol{v}^T$ denotes the transpose of vector $\boldsymbol{v}$.

## II. Methodology

The flowchart of our approach of clustering DH users is shown in Fig. 2. The whole procedure can be divided into five steps, i.e., data collection, data preprocessing, feature extraction, user clustering, and result interpretation.

*1) Data collection:* The heat usage of multiple DH users and utility production output for a relatively long period of time are recorded with sufficient precision by smart metering or other facilities. The local weather conditions are also collected.

*2) Data preprocessing:* Data preprocessing must be carried out on the raw data, before extracting meaningful pattern features. In this paper, data preprocessing considers three major steps. The first performs detection of missing data. Absent hourly readings are interpolated using neighbor values. However, if within a single day, a relatively large number of missing readings (e.g., larger than 4) occur together, the

data of the whole day are discarded since too much information may be lost. The second step carries out detection of anomalous data readings. Once an anomalous value is identified, it is replaced by interpolation.

The intention of this paper is to find DH users' consumption patterns and cluster them into different groups. If a DH user does not have notable consumption pattern, it does not contribute to the final clustering result. Even worse, its existence may blur the boundary of two distinct clusters. Therefore, the third step of preprocessing tries to locate users with almost constant consumptions and then discard them from further analysis. Furthermore, if a user has almost no consumption variation in certain days (e.g., in the summer time when space heating stops), the data in these days can also be discarded since they may lead to difficulties in discovering true patterns.

*3) Feature extraction:* A major part of this paper is to present three novel features that can represent heat consumption patterns. These features are extracted from the preprocessed dataset, and then serve as the input to the DH user clustering step. The details are elaborated in Section III.

*4) User clustering:* The extracted pattern features are respectively sent to a clustering algorithm to partition DH users into separate groups.

*5) Result interpretation:* DH companies can use the clustering results to attain the knowledge of their customers' usage behaviors. The user partitions can also be compared to show whether different features contain the similar information and result in similar clustering results.

## III. Extraction of Pattern Features

Following Section II, a case study is used to demonstrate the execution of the proposed DH user clustering approaches.

### *A. Dataset description and preprocessing*

Our dataset was provided by a Swedish DH company. The raw user consumption data contain hourly heat consumption readings (in MW) of 561 users collected from 1st January to 13th December 2010, resulting in a total of 8328 heat consumption records for each user. Apart from the consumption data, the function types of the users are also provided by the DH company. Among the 561 DH users, 233 are labeled as "Multi-family House", 89 as "Office and School", 116 as "Commercial", 33 as "Hospital and Social Service", 33 as "Industry", and 57 are labeled as "unknown" due to some reasons. It is easy to observe from the data that the same type of users can have very diverse consumption patterns, in terms of both short- and long-term variation tendency, but users of different types may have similar behaviors.

In addition, hourly recorded weather data of the city for the same period were collected. Attributes include outdoor temperature, wind speed, relative humidity, solar radiation, and so forth. In this paper only the temperature is utilized. The heat production of the DH company was also hourly recorded, which is used to identify network heat demand.

A careful data preprocessing procedure is carried out. No missing data are identified. But a number of possibly anomalous records are found. In our experiment, the local outlier factor (LOF) method [26] is employed to detect the anomalous values. This method stems from the fact that if the difference between a reading and its neighbors is significantly larger than others, this reading is likely to be an outlier. A larger value of LOF indicates more chance that the data point is an outlier. After being identified (and verified through manual checking), the anomalous values can be replaced by interpolation, e.g., the average of their neighbor values.

Further, the standard deviations of the consumption readings of 114 users are found to be below 0.01. Through manual checking, it is found that 81 among them had almost constant readings during the whole year, and thus did not exhibit any notable usage pattern. They are removed from our analysis in order to avoid affecting the clustering results. This results in a total of $480$ users to be clustered, which include 198 labeled as "Multi-family House", 79 as "Office and School", 96 as "Commercial", 26 as "Hospital and Social Service", 30 as "Industry", and 51 as "unknown".

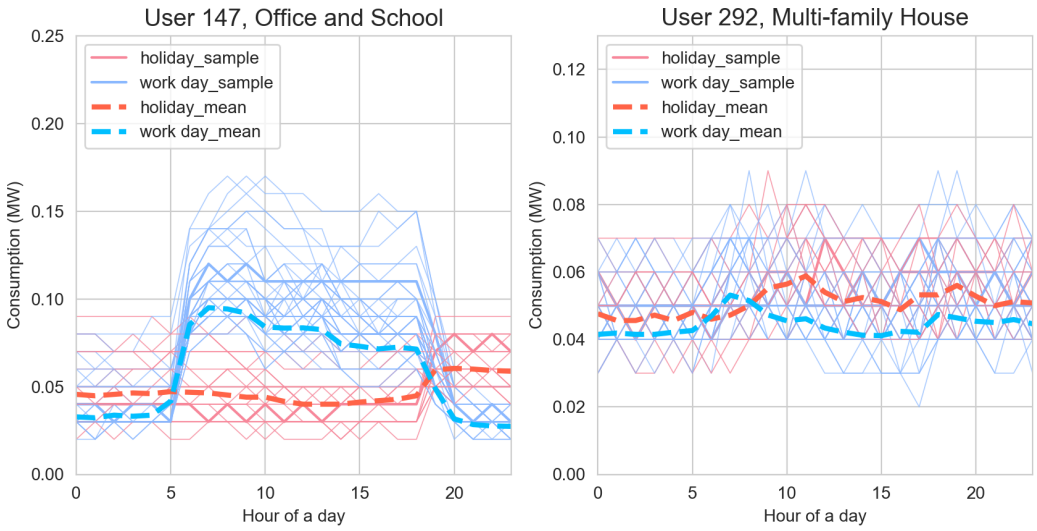



Fig. 3: Daily consumption variation curves of two users. Each light thin curve represents the daily consumption of a randomly sampled day. Thick dashed curves denote the average of these data.

Finally, the coefficient of variation (CV) is calculated, which is defined as the ratio between standard deviation and mean value, for each day of each user. The data of those days with CV smaller than $0.1$ are marked and discarded when conducting clustering using the features MDLP and CPC, since their almost constant daily usage behavior (mostly happened in summer time) may bias the clustering results.

### *B. Extraction of modified daily load profile (MDLP)*

Fig. 3 displays the daily consumption curves (light thin curves) of a few sampled winter days of two users in our dataset. Holiday and workday consumptions are shown by different colors. It can be clearly seen that each user has similar workday usage pattern. (The same observation also holds for the holiday consumption.) But the general variation tendencies of the two users are very different. Hence such a consumption pattern, represented by the average heat usage curve (thick dashed curves), can be used to facilitate clustering DH users. In addition, note that some users, such as User 147, can have very different usage behaviors in work days and holidays. Our first pattern feature, termed *modified daily load profile (MDLP)*, is generated to reflect the typical daily consumption variation tendency of each user. From Fig. 1(b) it can be seen that when the outdoor temperature was low, the user consumptions had a strong linear correlation with the temperature, mainly due to space heating. If the conventional approach is adopted to establish DLP by averaging all the daily usage curves of each user, all users tend to have the similar result since the DLP would be highly influenced by cold-day patterns and change as temperature changes. Therefore, in order to balance the impact of low- and high-temperature consumptions on the clustering result and better reveal the user-specific daily usage pattern, the DLP is modified by reducing the impact of temperature-related usage.

Specifically, for each day, the heat consumption is considered to consist of two main parts. The first corresponds to space heating. When outdoor temperature is lower than a certain threshold $T_{\text{sh}}$, a significant amount of heat usage, which is roughly linearly related to outdoor temperature, is consumed. The other part of heat usage is caused by regular user-specific activities, apart from space heating. The amount of the second part of consumption is assumed to be related to the time of the day and this usage behavior can be similar for different days. Characterizing this heat usage pattern is of interest.

To this end, each day is divided into $n$ time slots, each of which contains $\mu = \frac{24}{n}$ hours. Use a quantitative variable $H$ to denote sum heat consumption in one time slot, use a quantitative variable $T$

to denote the average outdoor temperature of that time slot, and use an $n$-level categorical variable $\boldsymbol{A}$ to denote the index of that time slot. When the temperature $T \leq T_{\text{sh}}$, the relationship between $H$, $T$, and $\boldsymbol{A}$ is modeled using a multiple linear regression model:

$$H = \beta_0^l + \beta_1^l(T_{\text{sh}} - T) + \boldsymbol{\beta}_2^l \boldsymbol{A} + \epsilon^l, \tag{1}$$

where the regression coefficient $\beta_1^l$ denotes the average consumption change when temperature $T$ changes by 1 degree, $\beta_0^l$ denotes the low-temperature model intercept, $\boldsymbol{\beta}_2^l$ quantifies the association between $\boldsymbol{A}$ and $H$, and $\epsilon^l$ is the error term that reflects the impact of other unknown factors.

When the outdoor temperature $T > T_{\text{sh}}$, space heating stops. As one can observe from Fig. 1(b) that in the high temperature region the linear dependency of consumption $H$ on outdoor temperature disappears, the relationship between $H$ and $\boldsymbol{A}$ is modeled using a simple linear regression model:

$$H = \beta_0^h + \boldsymbol{\beta}_2^h \boldsymbol{A} + \epsilon^h, \tag{2}$$

where $\beta_0^h$ denotes the high-temperature model intercept, $\boldsymbol{\beta}_2^h$ quantifies the association between $\boldsymbol{A}$ and $H$, and $\epsilon^h$ is the error term.

In order to incorporate categorical values into the linear regression models, $\boldsymbol{A}$ is represented using $n$ dummy binary variables: $\boldsymbol{A}$ is set as an $n$-dimensional vector $[\alpha_1, \alpha_2, ..., \alpha_n]^T$ where $\alpha_i \in \{0, 1\}$ and $\sum_{i=1}^{n} \alpha_i = 1$. The $i$th element $\alpha_i = 1$ means that the time index is $i$. Correspondingly, $\boldsymbol{\beta}_2^l = [\beta_{2,1}^l, \beta_{2,2}^l, \cdots, \beta_{2,n}^l]$ and $\boldsymbol{\beta}_2^h = [\beta_{2,1}^h, \beta_{2,2}^h, \cdots, \beta_{2,n}^h]$ are each taken as an $n$-dimensional vector, the $i$th element of which reflects the impact of user-specific behavior at time index $i$ on heat consumption.

For every user, there are a total of $2n + 4$ unknown coefficients in (1) and (2), $\beta_0^l$, $\beta_1^l$, $\beta_{2,1}^l$, $\cdots$, $\beta_{2,n}^l$, $\beta_0^h$, $\beta_{2,1}^h$, $\cdots$, $\beta_{2,n}^h$, and $T_{\text{sh}}$, to be learned from data. One way to solve this problem is to apply the Bayesian generalized linear model and use Markov chain Monte Carlo (MCMC) algorithms to find their posterior estimates (see, e.g., [27]). A more straightforward approach is to combine least squares with cross-validation to estimate the coefficients.

Specifically, similar to [28], workday and holiday patterns are distingished. Let us start from the work days. For each DH user two temperature values $T_{\text{sh},l}$ and $T_{\text{sh},u}$ are determined such that almost surely the true temperature threshold $T_{\text{sh}}$ for the user is bounded by them as $T_{\text{sh},l} < T_{\text{sh}} < T_{\text{sh},u}$ ($T_{\text{sh},l}$ and $T_{\text{sh},u}$ can be determined by simply visualizing the relationship between temperature and consumption of the user). In addition, define a sufficiently small step value $\triangle$. Now, for each value of $s \in \left\{1, 2, \cdots, \left\lfloor \frac{T_{\text{sh},u} - T_{\text{sh},l}}{\triangle} \right\rfloor \right\}$, set $T_{\text{sh}} = T_{\text{sh},l} + s\triangle$ and fit the models (1) and (2) with all workday consumption data and temperatures. Applying the least squares approach and 10-fold cross-validation [29], the index $s$ that leads to the smallest validation residual sum of squares (RSS) can be identified, and then the resulting temperature $T_{\text{sh},l} + s\triangle$ is chosen as the best estimate of workday temperature threshold for the user, denoted by $\hat{T}_{\text{sh}}$. The least squares estimates of the remaining regression coefficients when $T_{\text{sh}} = \hat{T}_{\text{sh}}$ are denoted by $\hat{\beta}_0^l$, $\hat{\beta}_1^l$, $\hat{\beta}_{2,1}^l$, $\cdots$, $\hat{\beta}_{2,n}^l$, $\hat{\beta}_0^h$, $\hat{\beta}_{2,1}^h$, $\cdots$, $\hat{\beta}_{2,n}^h$.

We intend to discard the heat consumption that is linear to outdoor temperature. To this end, when ambient temperature is lower than $\hat{T}_{\text{sh}}$, the value $\hat{\beta}_0^l + \hat{\beta}_{2,i}^l$ is used to represent the average consumption at time instant $i$ ($i \in \{1, 2, \cdots, n\}$) of each day without the linear impact of temperature. When temperature is higher than $\hat{T}_{\text{sh}}$, the consumption value is set to $\hat{\beta}_0^h + \hat{\beta}_{2,i}^h$. Then the typical workday consumption of a DH user at time instant $i$, denoted as $\hat{u}_i$, can be taken as a weighted sum of the low- and high-temperature consumptions to balance their influence:

$$\hat{u}_i = \hat{\gamma}\left(\hat{\beta}_0^l + \hat{\beta}_{2,i}^l\right) + (1 - \hat{\gamma})\left(\hat{\beta}_0^h + \hat{\beta}_{2,i}^h\right), \tag{3}$$

where $0 \leq \hat{\gamma} \leq 1$ denotes the weighting factor that can be set to represent the importance of low-temperature pattern. In our experiment $\hat{\gamma}$ is chosen to be the ratio of the number of low-temperature workday samples to the total number of workday samples of the associated user.

Similarly, following the above approach, fit the linear regression models (1) and (2) with holiday consumptions and temperatures, and obtain the least squares estimates of regression coefficients $\tilde{\beta}_0^l$, $\tilde{\beta}_1^l$,

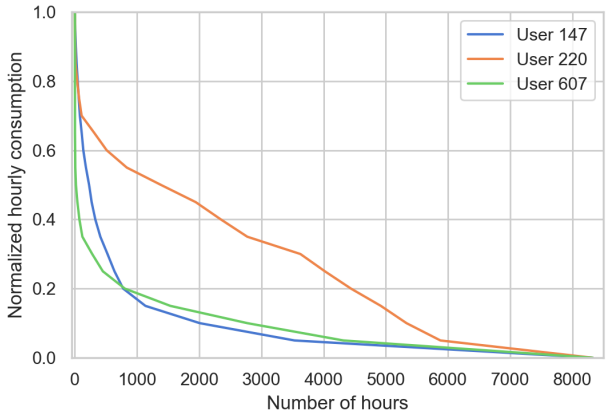



Fig. 4: Normalized duration curves of three example users.

$\tilde{\beta}^l_{2,1}, \cdots, \tilde{\beta}^l_{2,n}$, $\tilde{\beta}^h_0$, $\tilde{\beta}^h_{2,1}, \cdots, \tilde{\beta}^h_{2,n}$, and $\tilde{T}_{\text{sh}}$. The typical holiday consumption of a DH user at time instant $i$ can be taken as:

$$\tilde{u}_i = \tilde{\gamma}\left(\tilde{\beta}^l_0 + \tilde{\beta}^l_{2,i}\right) + (1-\tilde{\gamma})\left(\tilde{\beta}^h_0 + \tilde{\beta}^h_{2,n}\right). \tag{4}$$

where $0 \leq \tilde{\gamma} \leq 1$ is the weighting factor, taken as the ratio of the number of low-temperature holiday samples to the total number of holiday samples of the user in our experiment.

Let $\boldsymbol{u} = [\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_n, \tilde{u}_1, \tilde{u}_2, \cdots, \tilde{u}_n]$. Then $\boldsymbol{u}$ represents the daily consumption pattern of a user, when the usage linearly correlated to temperature is removed. In order to reflect only the variation tendency of each user's usage instead of absolute values, define the MDLP as the normalized version of $\boldsymbol{u}$, i.e.,

$$\boldsymbol{a} = \frac{\boldsymbol{u} - \text{mean}\{\boldsymbol{u}\}}{\text{std}\{\boldsymbol{u}\}}. \tag{5}$$

Now, $\boldsymbol{a}$ can be considered as a $2n$-dimensional feature vector for a user. Clustering using $\boldsymbol{a}$ can group DH users according to their typical daily consumption variation patterns.

### *C. Extraction of discretized duration curve (DDC)*

The second pattern feature is termed *discretized duration curve (DDC)*. It reflects whether a DH user has balanced usage of different consumption levels in a relatively long period of time. DDC is generated to represent the concept of duration curve, which is widely used in energy systems for economic dispatching and system planning [30]. It is able to reveal a long-term consumption variation tendency that cannot be shown by MDLP. In principle, duration curve is a plot of different consumption levels within a certain time period where the vertical axis represents consumption magnitude and the horizontal axis represents the length of time for which the usage exceeds that magnitude [30]. If a DH user has impulsive and significantly larger peak consumptions than its mean consumption, its duration curve would have a steep descending tendency. A relatively flat duration curve reflects a stable and balanced usage behavior throughout the considered time duration. In order to emphasize only the overall descending pattern instead of absolute values, the approach of plotting duration curve is modified by scaling (normalizing) the consumption of each user, denoted by vector $\boldsymbol{h}$, to between 0 and 1 as:

$$\bar{\boldsymbol{h}} = \frac{\boldsymbol{h} - \min\{\boldsymbol{h}\}}{\max\{\boldsymbol{h}\} - \min\{\boldsymbol{h}\}}, \tag{6}$$

The scaled duration curves of three example users in our dataset are displayed in Fig. 4. Different patterns can be clearly observed.

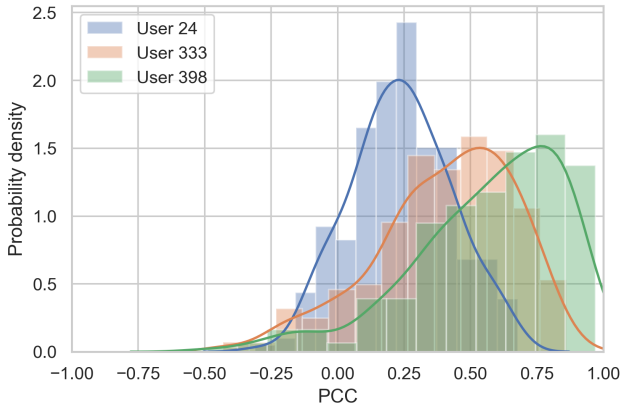



Fig. 5: Histogram of the Pearson correlation coefficients (PCC) of three example users. An estimated distribution curve for each histogram is also illustrated.

To describe the descending tendency with limited number of feature dimensions for facilitating DH user clustering, the normalized duration curves are discretized. Define the DDC of each user to be an $m$-dimensional vector $\boldsymbol{b} = [b_1, b_2, \cdots, b_m]$, where $b_i$ ($i \in \{1, 2, \cdots, m\}$) is the number of sample durations (i.e., hours) that the normalized consumption exceeds the value $\frac{i}{m+1}$. Clustering using $\boldsymbol{b}$ can group DH users according to their long-term consumption variation patterns.

### *D. Extraction of consumption-production consistency (CPC)*

Similar to MDLP, the third pattern feature, termed *consumption-production consistency (CPC)*, is derived based on the daily consumption behavior of each user. But instead of averaging all the daily consumptions to reach a single representative consumption variation pattern, it measures the similarity between the user's consumption and the DH company's heat production load. If high similarity frequently appears (we say that the consistency is high), it is more likely that the user has high contribution to the variation of network load demand and tends to cause peak loads.

For the $d$th day (recall that the days with almost constant consumption are not taken into consideration), let a DH user's daily consumption be denoted by a 24-level vector $\boldsymbol{h}_d$ and the DH company's production output by $\boldsymbol{g}_d$. The similarity between them is measured using the PCC [29]:

$$\rho_d = \frac{\text{cov}\{\boldsymbol{h}_d, \boldsymbol{g}_d\}}{\text{std}\{\boldsymbol{h}_d\} \times \text{std}\{\boldsymbol{g}_d\}}. \tag{7}$$

In principle, PCC measures the linear correlation between two random variables. It varies between $-1$ and 1, where the boundaries 1 and $-1$ respectively mean total positive and total negative linear correlations, and 0 implies no linear correlation at all.

Fig. 5 displays the (normalized) histogram of the 347 values of PCC for three example users in our dataset. It is seen that the DH users can have different distributions of the PCC. For example, User 398's daily consumption variation behaviors are more consistent with the DH company's production load pattern than the other two users, since its $\rho_d$ are more frequently close to 1. To compare different users' consumption-production consistency patterns, divide the region between $-1$ and 1 into a total of $l$ equal-size sub-regions. The CPC of each user is defined as an $l$-dimensional vector $\boldsymbol{c} = [c_1, c_2, \cdots, c_l]$ where $c_i$ is the occurrence frequency (ratio of the number of occurrence to the total number of samples) of the user's PCC in the $i$th sub-region, for $i \in \{1, 2, \cdots, l\}$. Clustering using $\boldsymbol{c}$ can group DH users according to the level of consistency between individual consumption and the network load demand seen by the DH company.

The next section provides the clustering results using these pattern features.

## IV. Results of Clustering DH Users

### A. Clustering algorithm and the number of clusters

A large number of clustering algorithms have been developed to tackle different problems from different perspectives (see, e.g., [31]). On the one hand, clustering is an unsupervised learning technique. Due to the lack of data labels, the research is exploratory by nature and it is normally hard to evaluate the optimality of applying a clustering algorithm on a dataset [29]. Since in general there is no ground truth of how DH users should be clustered, it is challenging to identify the optimal clustering algorithm for each feature to partition the users in the best way. On the other hand, it is widely known that as long as the features and distance measures are chosen properly with clear meaning, even a simple algorithm such as the $k$-means algorithm can also achieve relatively good performance [31]. The main focus of this paper is finding features that can represent the DH user consumption pattern from different points of view. Therefore, as a demonstration purpose, the GMM clustering algorithm [32] is applied to partition the DH users in our dataset. GMM is based on a probabilistic model which assumes data points to be generated from a mixture of $k$ (possibly multi-dimensional) Gaussian distributions. The parameters of the Gaussian distributions are iteratively optimized to better fit the data, and then the most likely cluster that each data point belongs to can be identified. In addition to the GMM, the $k$-means algorithm can also be applicable, as shown in [33].

A common challenge to all clustering techniques is how to optimally decide $k$, the number of clusters that the data should be partitioned into. An approach to address this problem is to try different choices of $k$ and treat finding the best value of $k$ as a model selection problem [31]. A number of performance indicators have been proposed to facilitate comparing the models. However, different indicators can often lead to different conclusions. Again, due to the lack of ground truth, it is hard to decide the optimal indicator. In practice, the best way of determining the proper number of DH user clusters should be combining mathematical models with domain experts'/utility companies' opinions. But this method is unavoidably subjective to a certain extent and the result would depend on the problem at hand. Thus this paper sticks to objective mathematical analysis. For the applied GMM clustering algorithm, the Bayes information criterion (BIC) [34] is used as the performance indicator for model selection.

Fitting data to a more complex model in general leads to better performance, but also potentially results in overfitting. To solve this problem, the BIC gives a higher penalty to models with more parameters. When applying BIC, for each choice of $k$, treat GMM as $k$ different Gaussian sub-models, denoted by $M_j$ ($j \in \{1, 2, \cdots, k\}$). The BIC for the $j$th sub-model is defined as

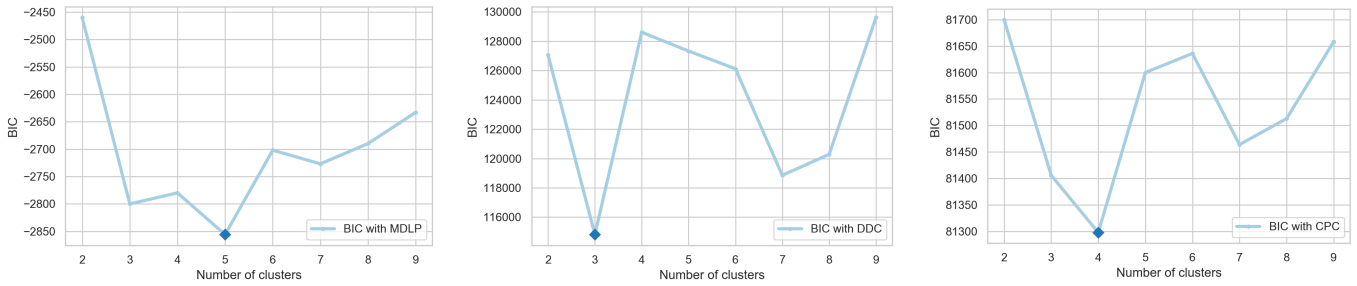
$$BIC_{M_j} = k_j \log N_j - 2 \log L_j, \tag{8}$$

where $k_j$ is the number of parameters in $M_j$, $L_j$ is the maximum value of the likelihood function $L_j = p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j)$, $\hat{\boldsymbol{\theta}}_j$ is the maximum likelihood estimation parameter vector, $\boldsymbol{x}$ is data, and $N_j$ is the number of data points in $M_j$, i.e. the number of users that belong to the $j$th cluster. The total BIC is $BIC = \sum_{j=1}^{k} BIC_{M_j}$. The best value of $k$ is determined when BIC achieves the smallest value.

### B. Clustering using MDLP

DH user clustering is first conducted using the pattern feature MDLP. In our experiment, the parameters are chosen to be $n = 12$ (i.e., the consumptions of two hours are combined as a single value), $T_{\text{sh},l} = 10°\text{C}$, $T_{\text{sh},h} = 20°\text{C}$, and $\triangle = 0.2°\text{C}$. By this means, the MDLP of each user, $\boldsymbol{a}$, is a $24$-dimensional vector. The first and remaining $12$ elements in $\boldsymbol{a}$ respectively represent the (bi-hourly sampled) workday and holiday consumption variation patterns, where the consumptions linearly related to low outdoor temperature are discarded.

The obtained MDLP feature vectors of the 480 users are sent to the GMM clustering algorithm. Fig. 6(a) displays the BIC for different choices of the number of clusters $k$. It is seen that $k = 5$ leads to the smallest BIC and thus in this paper the 480 DH users are partitioned into $5$ clusters. However, as mentioned earlier,




(a) (b) (c)

Fig. 6: Bayes information criterion for clustering using (a) modified daily load profile, (b) discretized duration curve, and (c) consumption-production consistency.

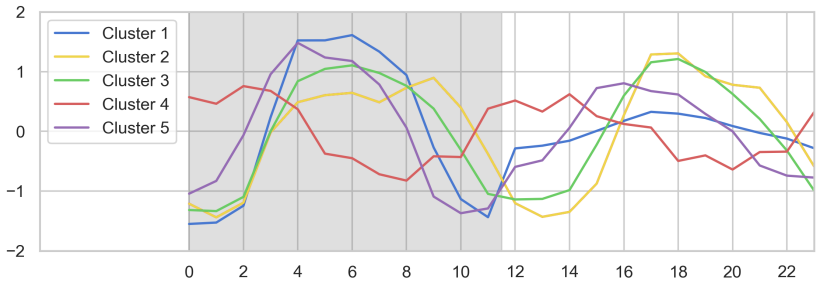



Fig. 7: Cluster centroids for clustering using modified daily load profile with $k = 5$. Workday patterns are shown in the gray area, and holiday patterns in the white area.

using other clustering methods and performance indicators, or even having another dataset, may lead to different results. The practically optimal solution may be combining such analysis with the opinions of domain experts/utility companies to ensure the clustering results to have more meaningful interpretations.

Based on the BIC, set $k = 5$ and plot the cluster centroids (the 24 elements are connected for ease of illustration) in Fig. 7. Fig. 8 plots individual user's (connected) MDLP features. The distributions of user function types and the number of users placed in each cluster are also shown. It is observed that the users initially labeled by their functions cannot be well separated by usage pattern clustering. This confirms that the same type of users do not necessarily have the same heat usage behavior. In the figures, cluster 1 shows a pattern with much higher usage in work days than that in holidays. The former changes significantly over 24 hours, but the latter is relatively stable. Cluster 2 shows a pattern with two peaks during work days (the major fraction of users are "Multi-family House"). Clusters 3 and 5 include the users whose workday and holiday patterns are similar, but the typical daily usage variation of the former lags behind that of the latter. Cluster 4 consists of users whose pattern cannot be included in other four groups. These results can be sent for interpretation from domain experts' viewpoints for developing energy reduction and pricing strategies.

### C. Clustering using DDC

Now, apply the GMM algorithm using the pattern feature DDC generated from our dataset. As a demonstrative example, set $m = 24$, i.e., the extracted DDC feature vector $\boldsymbol{b}$ also has 24 elements. Fig. 6(b) shows that $k = 3$ leads to a better BIC score than other choices. Again, it is recommended that the opinions of domain experts and utility companies should be taken into account in practice.

Fig. 9 illustrates the (connected) cluster centroids for choosing $k = 3$. The associated distribution of user function labels and the number of users in each cluster are also shown. Again, users of the same function type cannot be well grouped. This means they have different patterns. As mentioned in Section

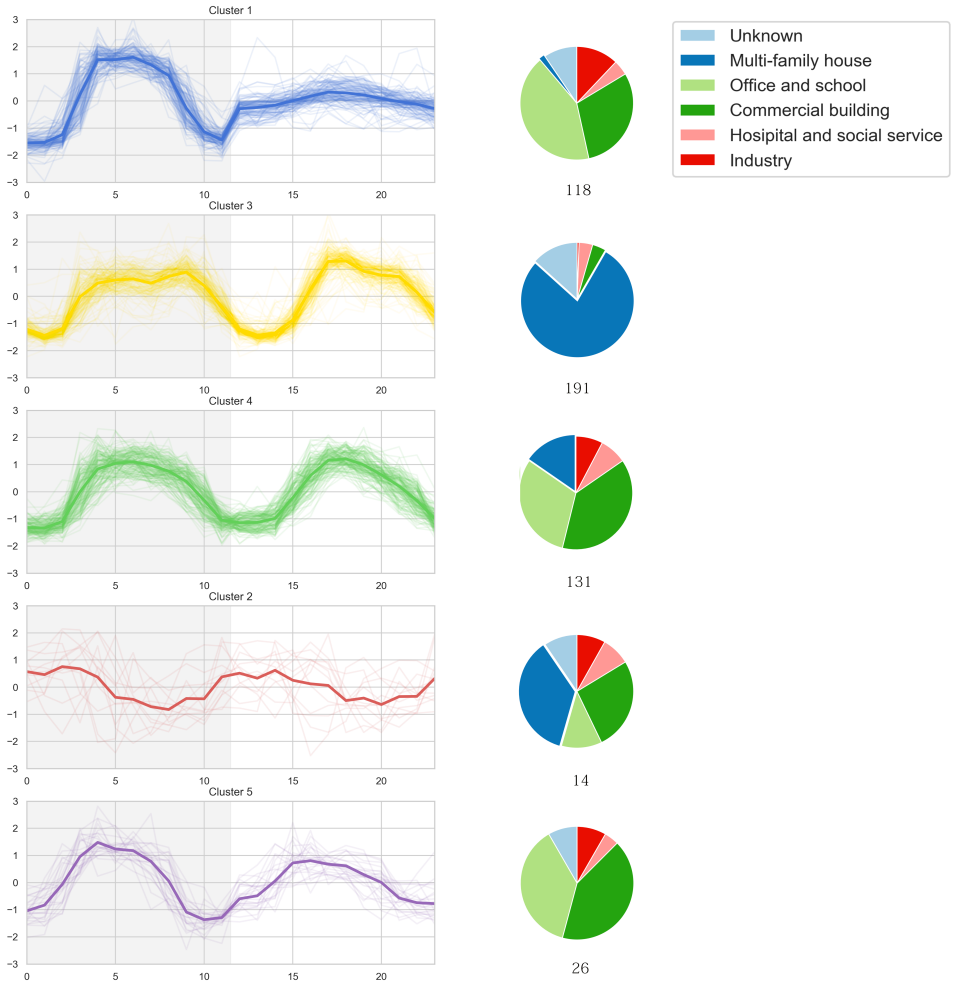



Fig. 8: Individual user patterns and user distribution for clustering using modified daily load profile with $k = 5$. User patterns are shown as light curves. Each cluster centroid is shown as a solid curve. User function type distributions are shown in pie charts. The number under each pie chart displays the number of users in that cluster.

III-C, a steep descending pattern of the normalized duration curve implies that the user has impulsive peak consumption much larger than its mean consumption. Hence the long-term usage behavior is relatively unstable and unbalanced. DH users grouped into clusters with such cluster centroids, e.g., Cluster 3 in Fig. 9, may thus deserve more attention.

### *D. Clustering using CPC*

Finally, following Section III-D, CPC is extracted from the user consumption data and utility production data. Again, choose $l = 24$. The CPC feature vector for each user is a 24-dimensional vector $\mathbf{c}$. The BIC displayed in Fig. 6(c) shows that $k = 4$ can be adopted as the number of clusters.

Fig. 10 depicts the (connected) cluster centroids, the associated user type distribution, and the number of users placed in each cluster. Again, the users of the same type can have very different usage patterns from the viewpoint of CPC. The clustering result in Fig. 10 reveals four groups of users which cause different levels of pressure on the utility production capacity. Compared with those in other groups, the users in Cluster 4 frequently have daily usage variation similar to that of the utility production load, and thus are more likely to have high contribution to network peak load.

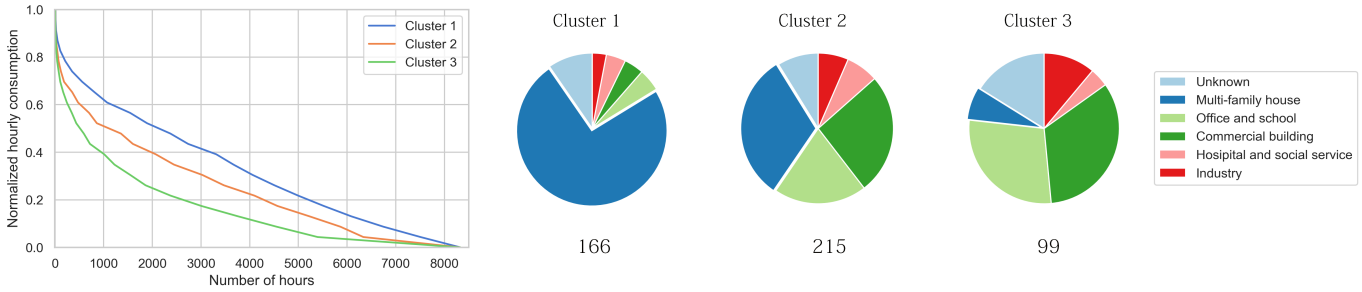



Fig. 9: Cluster centroids and user function type distributions for clustering using discretized duration curve with $k = 3$.

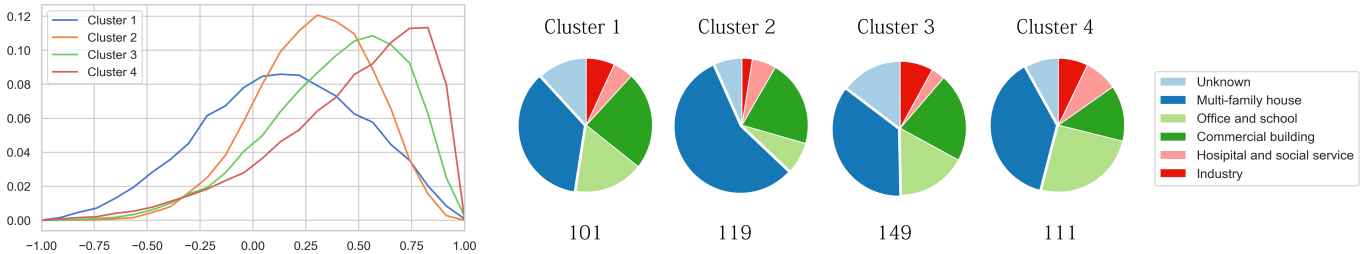



Fig. 10: Cluster centroids and user function type distributions for clustering using consumption-production consistency with $k = 4$.

### *E. Comparison of clustering results*

So far, three different clustering results are obtained using three pattern features. DH companies can use these results to separate their users into groups from different perspectives. How different these clustering results are can also be measured. Anyhow, if two features lead to almost the same way of partitioning the DH users, they may imply the similar information and hence one of them would be redundant. To this end, one can follow [35] and apply the concept of normalized mutual information (NMI) to evaluate the similarity of any two clustering results.

In information theory, the entropy of a discrete random variable $X$ with alphabet $\mathcal{C}$ and probability mass function (PMF) $p(x)$ is defined as [36]:

$$H(X) = -\sum_{x \in \mathcal{C}} p(x) \log p(x). \tag{9}$$

It represents the average uncertainty contained in $X$, i.e., the average information obtained (uncertainty removed) by observing $X$. The mutual information between $X$ and another random variable $X'$ with alphabet $\mathcal{C}'$ and PMF $p(x')$ is

$$I(X; X') = -\sum_{x \in \mathcal{C}, x' \in \mathcal{C}'} p(x, x') \log \frac{p(x, x')}{p(x)p(x')}, \tag{10}$$

where $p(x, x')$ is the joint PMF. Essentially $I(X; X')$ quantifies the reduced uncertainty in $X$ by observing $X'$. Hence it measures the mutual dependence between $X$ and $X'$. The fact that $I(X; X')$ is not bounded makes it difficult to interpret when $I(X; X')$ is used to measure the similarity of clustering results. Hence the NMI is adopted in [37]:

$$\bar{I}(X; X') = \frac{2I(X; X')}{H(X) + H(X')}. \tag{11}$$

The NMI is a metric, and varies between 0 and 1, where $\bar{I}(X; X') = 1$ suggests that $X$ and $X'$ can fully explain each other, and $\bar{I}(X; X') = 0$ indicates independence.

TABLE I: Normalized mutual information between clustering results.

| ＼ | MDLP | DDC | CPC |
|---|---|---|---|
| MDLP | 1.000 | 0.141 | 0.017 |
| DDC | 0.141 | 1.000 | 0.035 |
| CPC | 0.017 | 0.035 | 1.000 |

To apply the NMI to evaluate two partition results attained by two different features, the clustering result of each DH user is treated as a random variable, i.e., if one randomly chooses a user, to which cluster it belongs can be modeled as a random variable $X$. If one clustering algorithm partitions all DH users into $k$ clusters, denoted as $C_1, C_2, \cdots, C_k$, then $X$ has alphabet $\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$ and PMF $p(X = C_j) = \frac{|C_j|}{N}$ for $j \in \{1, 2, \cdots, k\}$, in which $|C_j|$ is the size of cluster $C_j$ and $N$ is the total number of users. Similarly, if another clustering algorithm partitions the DH users into $k'$ clusters, denoted as $C'_1, C'_2, \cdots, C'_{k'}$, the associated random variable, denoted as $X'$, has alphabet $\mathcal{C} = \{C'_1, C'_2, \cdots, C'_{k'}\}$ and PMF $p(X' = C'_j) = \frac{|C'_j|}{N}$ for $j \in \{1, 2, \cdots, k'\}$. The joint PMF $p(X = C_j, X' = C'_{j'})$ represents the probability that a user is set to $C_j$ and $C'_{j'}$ by the two algorithms respectively:

$$p(X = C_j, X' = C'_{j'}) = \frac{|C_j \cap C'_{j'}|}{N}. \tag{12}$$

Substituting these PMF expressions into (11) leads to the NMI.

The NMI between each pair of clustering results is shown in Table I. It can be seen that all results are relatively small, which indicates that the three proposed pattern features contain quite different amount of information. The clustering methods presented in this paper can hence lead to insights about DH users' consumption behaviors from different perspectives.

## V. Conclusion

Heat usage has its own characteristics. The consumption patterns of district heating (DH) users may not be revealed by user function labels or conventional user clustering approaches for electricity usage. In this paper, new methods based on three pattern features have been proposed to cluster DH users. The findings can provide new knowledge regarding DH consumption patterns to potentially benefit the development of better energy management and usage strategies.

- The ambient temperature has strong impacts on the heat demand of almost all users, which can conceal the influence of individual user's behaviors. In order to remove the consumption that is linearly related to ambient temperature, the modified daily load profile can be used, based on which the clustering results are able to better reflect user-specific consumption patterns.
- The discretized duration curve can be used to group DH users according to how balanced their consumptions are in a relatively long period of time. The clustering result can reveal long-term consumption variation tendencies that cannot be shown by MDLP.
- The consumption-production consistency can be used to reflect the similarity level between a DH user's consumption and the DH company's production load. The clustering result may help identify the users that are more likely to have high contribution to network peak load.

## Acknowledgements

Chao Wang would like to acknowledge the funding support of the National Natural Science Foundation of China (no. 61771343) and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant (no. 752979). This work reflects only the authors' view and the EU Commission is not responsible for any use that may be made of the information it contains.

# REFERENCES


[1] D. Ürge-Vorsatz, L. D. D. Harvey, S. Mirasgedis, and M. D. Levine, "Mitigating $CO_2$ emissions from energy use in the world's buildings," *Building Research & Information*, vol. 35, no. 4, pp. 379–398, 2011.

[2] International Energy Agency, "Energy efficiency," https://webstore.iea.org/market-report-series-energy-efficiency-2018, 2018.

[3] R. J. de Dear, T. Akimoto, E. A. Arens, G. Brager, C. Candido, K. W. D. Cheong, B. Li, N. Nishihara, S. C. Sekhar, S. Tanabe, J.Toftum, H. Zhang, and Y. Zhu, "Progress in thermal comfort research over the last twenty years," *Indoor Air*, vol. 23, no. 6, pp. 442–461, 2013.

[4] Z. Ma, J. Xie, H. Li, Q. Sun, Z. Si, J. Zhang, and J. Guo, "The role of data analysis in the development of intelligent energy networks," *IEEE Network*, vol. 31, no. 5, pp. 88–95, 2017.

[5] S. Ramos, J. M. M. Duarte, J. Soares, Z. Vale, and F. J. Duarte, "Typical load profiles in the smart grid context: a clustering methods comparison," in *IEEE Power and Energy Society General Meeting*, San Diego, USA, 22-26 Jul. 2012.

[6] Q. Sun, H. Li, Z. Ma, C. Wang, J. Campillo, Q. Zhang, F. Wallin, and J. Guo., "A comprehensive review of smart meters in intelligent energy networks," *IEEE Internet of Things Journal*, vol. 3, no. 4, pp. 464–479, 2016.

[7] Y. I. Kim, J. M. Ko, and S. H. Choi, "Methods for generating TLPs (typical load profiles) for smart grid-based energy programs," in *IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG'11)*, Paris, France, 11-15 Apr. 2011.

[8] V. Ford and A. Siraj, "Clustering of smart meter data for disaggregation," in *Global Conference on Signal & Information Processing (GlobelSIP)*, Atlanta, USA, 3-5 Dec. 2014.

[9] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, "Clustering analysis of residential electricity demand profiles," *Applied Energy*, vol. 135, pp. 461–471, 2014.

[10] F. L. Quilumba, W. J. Lee, H. Huang, D. Y. Wang, and R. Szabados, "An overview of AMI data preprocessing to enhance the performance of load forecasting," in *IEEE Industry Applications Society Meeting*, Vancouver, Canada, 5-9 Oct. 2014.

[11] A. Shahzadeh, A. Khosravi, and S. Nahavandi, "Improving load forecast accuracy by clustering consumers using smart meter data," in *International Joint Conference on Neural Networks (IJCNN'15)*, Killarney, Ireland, 12-17 Jul. 2015.

[12] H. Bagge and D. Johansson, "Measurements of household electricity and domestic hot water use in dwellings and the effect of different monitoring time resolution," *Energy*, vol. 36, no. 5, pp. 2943–2951, 2011.

[13] R. Granell, C. J. Axon, and D. C. H. Wallom, "Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3217–3224, 2015.

[14] F. Mcloughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Applied Energy*, vol. 141, pp. 190–199, 2015.

[15] D. Vercamer, B. Steurtewagen, D. V. D. Poel, and F. Vermeulen, "Predicting consumer load profiles using commercial and open data," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3693–3701, 2016.

[16] I. P. Panapakidis, "Clustering based day-ahead and hour-ahead bus load forecasting models," *International Journal of Electrical Power & Energy Systems*, vol. 80, pp. 171–178, 2016.

[17] I. P. Panapakidis, S. I. Frantza, and G. K. Papagiannis, "Implementation of price-based demand response programs through a load pattern clustering process," in *Mediterranean Conference on Power Generation (Medpower'16)*, Belgrade, Serbia, 6-9 Nov. 2016.

[18] H. Niska, "Extracting controllable heating loads from aggregated smart meter data using clustering and predictive modelling," in *IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP'13)*, Melbourne, Australia, 2-5 Apr. 2013.

[19] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer, "Cluster-based aggregate forecasting for residential electricity demand using smart meter data," in *IEEE International Conference on Big Data*, Santa Clara, USA, 29 Oct-1 Nov. 2015.

[20] Z. Ma, H. Li, Q. Sun, C. Wang, A. Yan, and F. Starfelt, "Statistical analysis of energy consumption patterns on the heat demand of buildings in district heating systems," *Energy & Buildings*, vol. 85, pp. 464–472, 2014.

[21] J. Xie, H. Li, Z. Ma, Q. Sun, F. Wallin, Z. Si, and J. Guo, "Analysis of key factors in heat demand prediction with neural networks," *Energy Procedia*, vol. 105, pp. 2965–2970, 2017.

[22] D. Yu, "A two-step approach to forecasting city-wide building energy demand," *Energy and Buildings*, vol. 160, pp. 1–9, 2018.

[23] J. Zhao, Y. Xin, and D. Tong, "Energy consumption quota of public buildings based on statistical analysis," *Energy Policy*, vol. 43, no. 2, pp. 362–370, 2012.

[24] C. E. Kontokosta and C. Tull, "A data-driven predictive model of city-scale energy use in buildings," *Applied energy*, vol. 197, pp. 303–317, 2017.

[25] J. Song, F. Wallin, and H. Li, "District heating cost fluctuation caused by price model shift," *Applied Energy*, vol. 194, pp. 715–724, 2017.

[26] C. C. Aggarwal, *Outlier Analysis*. Springer, 2017.

[27] J. Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[28] C. Chelmis, J. Kolte, and V. K. Prasanna, "Big data analytics for demand response: Clustering over space and time," in *IEEE International Conference on Big Data*, Santa Clara, USA, 29 Oct-1 Nov. 2015.

[29] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An Introduction to Statistical Learning*. Springer, 2016.

[30] J. Valenzuela and M. Mazumdar, "A probability model for the electricity price duration curve under an oligopoly market," *IEEE Transactions on Power Systems*, vol. 20, no. 3, pp. 1250–1256, 2005.

[31] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, pp. 651–666, 2010.

[32] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic Press, 2009.

[33] Y. Du, C. Wang, H. Li, J. Song, and B. Li, "Clustering heat users based on consumption data," in *International Conference on Applied Energy (ICAE'18)*, Hong Kong, China, 22-25 Aug. 2018.

[34] D. Pelleg and A. Moore, "X-means: Extended k-means with an efficient estimation of the number of clusters," in *Intelligent Data Engineering and Automated Learning (IDEAL'00)*, Hong Kong, China, 13-15 Dec. 2000.

[35] M. Meila, "Comparing clusterings—an information based distance," *Journal of Multivariate Analysis*, vol. 98, no. 5, pp. 873–895, 2007.
[36] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley, 2003.
[37] A. Lancichinetti, S. Fortunato, and J. Kertesz, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, pp. 19–44, 2008.