Clustering-based analysis for residential district heating data

Gianniou, Panagiota; Liu, Xiufeng; Heller, Alfred; Nielsen, Per Sieverts; Rode, Carsten

Published in:
Energy Conversion and Management

Link to article, DOI:
10.1016/j.enconman.2018.03.015

Publication date:
2018

Document Version
Peer reviewed version

Clustering-based Analysis for Residential District Heating Data

Panagiota Gianniou, Xiufeng Liu*, Alfred Heller, Per Sieverts Nielsen, Carsten Rode

Technical University of Denmark

Abstract

The wide use of smart meters enables collection of a large amount of fine-granular time series, which can be used to improve the understanding of consumption behavior and used for consumption optimization. This paper presents a clustering-based knowledge discovery in databases method to analyze residential heating consumption data and evaluate information included in national building databases. The proposed method uses the K-means algorithm to segment consumption groups based on consumption intensity and representative patterns and ranks the groups according to daily consumption. This paper also examines the correlation between energy intensity and the characteristics of buildings and occupants, load profiles of households, consumption behavior changes over time, and consumption variability. The results show that the majority of the customers can be represented by fairly constant load profiles. Calendar context has an impact not only on the patterns but also on the consumption intensity and user behaviors. The variability studies show that consumption patterns are serially correlated, the customers with high energy consumption have lower variability, and the consumption is more stable over time. These findings will be valuable for district heating utilities and energy planners to optimize their operations, design demand-side management strategies, and develop targeting energy-efficiency programs or policies.

Keywords: Clustering, Load pattern, Load profiling, Load transition, Variability

1. Introduction

Information and communications technologies (ICTs) are revolutionizing today's energy management systems. The distinct characteristic is to use smart meters to monitor energy consumption. Smart meters are the digital devices that can collect energy consumption at a fine-granular time level, typically every 15 minutes [1]. Smart metering systems are being widely installed globally. European countries have set the goal of converting their legacy meters to smart meters by 2020 in line with the Third Energy Package in the Electricity Directive [2] and Gas Directive [3] issued by European Commission in 2006. It is expected that 72% of European consumers will own electricity smart meters by 2020 and 40% of them will have gas meters [4]. In Denmark, more than half of current electricity customers have smart meters [5]. In recent years, several large-scale smart meter installation projects have been carried out, including the project undertaken by Enel SpA in Italy covering 30 million customers between 2000 and 2005, the Linky pilot project in France involving 300,000 customers and the national Australian initiative in Victoria covering a total of 2.6 million electricity customers. Furthermore, smart meters have been installed by Chubu Electric Power in Japan for most of their high-voltage customers, office buildings, and individual households.

The main drivers behind the employment of smart metering in different countries include load management, peak or demand reduction, fraud reduction, accurate billing and water conservation [6]. Smart meters can also be utilized to develop more accurate prediction models and detailed analyses on the drivers of building energy consumption [9]. For example, the data can reveal potentially valuable information about buildings and end users that are useful to energy management. In particular, building-related data can unveil hidden correlations between energy use and its influencing factors in buildings. Smart meter data can also help developing and applying control strategies to improve building energy performance and efficiency [10]. In addition, building energy-related information can be provided to customers [11], which can help them use demand-response techniques to reduce energy consumption, improve energy efficiency, reduce carbon emissions and improve the use of renewable energy sources. There is also an expectation that smart metering or advanced metering infrastructure can contribute to demand and cost reduction and to the adoption of domestic low-and zero-carbon technologies [6]. Therefore, they can be utilized to decrease uncertainty related to building energy performance and provide detailed information on energy monitoring. Smart meter data analyses are therefore not only of value to the research community and customers but they also help utilities to improve their meter-to-cash processes. They enable them to use advanced tariff schemes and ultimately contribute to the value of metering infrastructure [7, 8].

Buildings account for nearly one-third of the world's energy consumption. Among others, half of the global energy consumption of buildings comes from space heating and cooling as well as hot water [12]. In heating-dominated climates, such

*Corresponding author

Email address: xiuli@dtu.dk (Xiufeng Liu*)

as Scandinavia, the main source of building energy demand is space heating. The increasing availability of smart meter data makes it possible to gain insights into heating consumption of buildings to help with energy management, such as extracting hidden temporal patterns - knowledge which can not be captured at its detailed level without the use of smart meters. Despite the promising benefits, it is still difficult to obtain reliable and detailed heating data, largely due to commercial sensitivity and privacy issues. Moreover, heating sector management is typically more challenging than other energy sectors, such as electricity, due to the high variation of production and demand. It is therefore difficult to obtain reliable data on heat production, usage patterns, and production costs. In addition, the installation of heating meters is more challenging and costly than smart electricity meters [13]. This also means that data analyses in the heating sector receive much less attention.

In this paper, a clustering-based knowledge discovery in database approach is proposed for analyzing residential heating consumption data. The objective of this study is to provide the information to district heating utilities, which they can use for optimizing their operations and for better understanding their customers. At the same time, the analysis can be used by customers to understand their consumption profiles and behaviors to improve energy efficiency. In addition, this study can help utilities and decision makers to develop energy efficiency strategies and policies, as well as provide personalized energy services to specific customer segments. Due to the introduction of renewable energy sources and the electrification of the heating sector, the Danish energy sector is in a transition period, which means that the balancing of the power grid is challenging. Energy flexibility solutions can support this transition, including demand-side management techniques that balance utility and customer needs. Identification and mapping of consumption patterns enable district heating utilities to implement new operational strategies. This study will perform a statistical clustering analysis on heating consumption data of Danish dwellings connected to a local district heating network. The data consists of the load profiles of 8,293 single-family households from Aarhus, Denmark. The study clusters the customers into different groups by the K-means clustering algorithm, ranks the groups according to their consumption intensity, and labels them using different alphabets. An exploratory analysis is conducted on the energy consumption and socio-technical data of the dwellings, which reveals the correlation between energy consumption and the characteristics of the buildings and occupants.

This paper makes the following contributions: First, this paper presents a clustering-based approach for district heating consumption data analysis, including data preparation, clustering, and the analysis based on the clustering results. Second, this paper uses the K-means algorithm to segment consumption intensity groups, studies the correlation between the consumption intensity and the characteristics of buildings and occupants, and analyzes the transition of consumption behaviors over time. Third, this paper identifies representative patterns by clustering normalized daily consumption patterns and studies the variability of consumption patterns using an entropy approach. Fourth, this paper leads to a number of interesting findings from the clustering-based analysis including non-intuitive results, such as the calendar impact on consumption, the serial correlation of consumption patterns, and high consumption with a lower variability.

The remainder of this paper is organized as follows. Section 2 gives a review of related work. Section 3 presents the methods applied. Section 4 presents the data and the results of the clustering-based analysis. Section 5 discusses the significant findings of this paper and the related issues. Section 6 concludes the paper and points to the future research direction.

2. Literature review

Clustering is one of the most commonly used techniques in data mining. It is used to discover groups and identify interesting distributions in the underlying data [15]. Specifically, a clustering problem is about partitioning a given data set into groups, classes or clusters such that the data points in the same cluster are more similar compared with the data points in other clusters [15]. The goals of cluster analysis are data reduction, hypothesis development and testing and prediction identification, based on groups [16]. The main steps of clustering include feature selection, choice of the clustering algorithm, validation and interpretation of results [17]. Clustering algorithms fall into the following categories based on how clusters are defined: partitional clustering, including K-means and K-medoids methods, and hierarchical clustering, which includes density-based and grid-based clustering [18]. Clustering has been applied to different fields such as biology, web mining etc. Clustering techniques are being applied to the analysis of smart meter data in buildings and district heating systems, because of the rapid introduction of smart meters. Knowledge about the characteristics of clusters of consumers with similar consumption patterns can facilitate the development of novel tariff schemes and improve network management [19].

Most of the studies on clustering analysis using smart meter data are found in the electricity sector where smart meters are common [20]. In [21], the clustering methods for electricity consumption pattern recognition were assessed, including hierarchical and K-means algorithms. It was found that the K-means algorithm is more useful for segmenting customer groups, and that information hidden in the consumption patterns has a great potential to be exploited towards load control, demand response actions, real-time pricing etc. In [22], hourly electricity consumption data from 4,500 smart meters, covering all categories of Danish customers were used to identify the load profiles and classify them. It was concluded that modeling of individual load profiles should be differentiated between different categories of customers, such as between industrial and residential uses, between weekday and weekend, and between summer and winter. In [19], an automatic classification method was proposed based on self-organizing maps, which identified a set of household properties that could be deduced from electricity consumption data, such as the size of household and the income of occupants. Two methods were introduced in [23] for modeling total energy consumption of buildings with regards to

predictive accuracy and cluster stability: clusterwise regression and cluster validation methods to measure stability. Clusterwise regression gave very accurate predictions but relatively unstable clusters, while K-means gave more stable clusters but poor predictions in several clusters. Thus, clustering methods should be carefully selected considering the main objective, whether it is accuracy or stability. In [24], a Bayesian non-parametric clustering algorithm was used to distinguish specific features between electricity consumption patterns of 219 households in UK and Bulgaria. The main advantage of the method was that it was not necessary to pre-define the number of clusters before analyzing. However, the algorithm was a bit slower in data processing than other clustering techniques. It was found that features such as nationality, family size, and type of dwelling can be successfully assigned to the members of a specific cluster.

Fewer studies have been found on data analysis in the heating sector compared with the electricity sector. In [25], the heating data from 139 single-family houses in Denmark were investigated for identifying patterns in space heating profiles using the K-means algorithm. It was found that the heating load profiles varied with external load conditions (high, medium and low demand periods). Two main clusters of load profiles were identified for weekdays and weekend days, one of which was quite stable and the other had a higher variation throughout the day. The latter was characterized by two peaks, one in the morning and the other in the afternoon/evening. It was concluded that building characteristics like floor area, building year and type of space heating distribution system, the existence of children or teenagers and the postcode proved to explain the difference between the two patterns. In another study [26], typical daily heating profiles were identified in a three years data set for 19 high education buildings in Norway using a Partitioning Around Medoids clustering algorithm. The Pearson Correlation Coefficient was used to determine the dissimilarity measure to cluster the daily load profiles based on the similarity of the variability, rather than the magnitude similarity. The typical heating load profiles provided information on the peaks and troughs of the daily heating consumption, daily high heating consumption period and daily load variation. Two statistical analysis approaches were introduced in [27] to develop random domestic profiles based on 15-min interval time series data from 25 typical Danish households. The results showed that space heating consumption had larger variability during a year, compared with domestic hot water consumption, which was more sporadic and transient.

Many studies identify the shape of load profiles as the primary objective and determine stable representative clusters. These methods are commonly referred to as shape-based. A shape-based approach to classifying energy consumption profiles at the household level was presented in [28]. The method was based on a dynamic time warping method that reflected the effect of hidden patterns of regular end-user behaviors. This method resulted in a smaller number of clusters, higher clustering efficiency and a lower household variability. In another study [29], a k-shaped approach was introduced, relying on a scalable iterative refinement process that created homogeneous and well-separated clusters. In order to consider the shape of the time series, a normalized version of the cross-correlation measure was used as the distance metric for the clustering. It was concluded that the proposed shape-based approach outperformed all scalable and non-scalable approaches in terms of accuracy, it was domain-independent and highly efficient for time series analysis.

In summary, clustering algorithms have been widely used for customer segmentation, pattern recognition, and classification. Most of the literature uses electricity data, which are more readily available and accessible. There is much less research carried out for residential heating consumption data analysis. Load shape recognition is the major objective of the reviewed articles. In contrast, in this paper, the analysis is conducted based on the clustering results, including load profiling, consumption behavior change over time, and the variability of the heating consumption patterns.

3. Methods

This paper uses a three-stage approach to process and analyze district heating consumption data. Figure 1 describes an overview of this approach, which includes data preparation, clustering and the analysis based on the clustering results. First, the heating consumption data are collected by the preparation module. The data usually contain some anomalous and missing consumption values that need to be cleaned. Data cleansing is an important step for ensuring data quality for the subsequent clustering-based data analysis. The data are pre-processed by filtering the single-family household data, removing the data with missing values, and adjusting the reading time stamp to align winter and summer time. The second module does the clustering on the load profiles. The K-means algorithm is employed for the clustering to segment different customer groups for further analysis in the third module. Two clustering analyses are applied. The first one is applied on the non-normalized data to identify the consumption intensity groups. The second clustering analysis is applied on the normalized data to identify the representative load patterns of all customers. It is expected that the first clustering analysis highlights the difference in the magnitude of heating consumption among customers. The second analysis points out the patterns of heating loads with regards to time. The optimal number of clusters is determined separately in both approaches, and different labels are assigned to the identified customer groups. The analysis of the clustering results is conducted in the third module, including customer segmentation according to the clustering results, logistic regression analysis to identify influencing factors on heating consumption, load profiling for individual customers, studying load transition over time, and consumption variability. The analysis algorithms are implemented using the in-database machine learning library, Apache MADlib [30], and run as database procedures in the open source database management system, PostgreSQL [31].

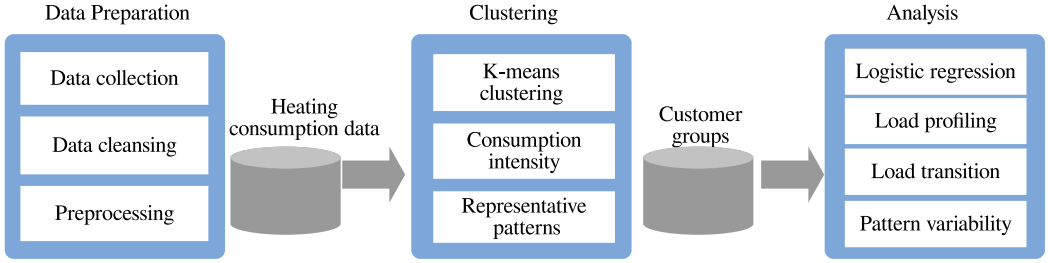

Figure 1: Overview of the clustering-based analysis

3.1. Clustering

The K-means clustering algorithm [32] is used in this analysis for clustering district heating consumption data. The goal of the clustering is to identify and segment the customers with similar load intensity and consumption patterns in the first place and secondly to specifically look at the consumption patterns on normalised data. Given a set of load patterns (daily load patterns based on hourly readings in this paper), the clustering algorithm classifies the load patterns into K groups or clusters. The patterns are similar within the same cluster, but dissimilar to the other clusters, according to a distance metric. K-means clustering is an iterative process with the aim of minimizing the intra-cluster inertia criterion defined by:

$$C(P,\mu) = \sum_{i=1}^{n} \sum_{X_i \in P_k} \|X_i - \mu_k\|^2 \tag{1}$$

where $P = (P_1, P_2, ..., P_k)$ is the set of clusters, $\mu = (\mu_1, \mu_2, ..., \mu_k)$ is the set of cluster centers, and $\| \cdot \|$ is the L_2 norm associated to the distance metric.

In this paper, the clustering approach is used for identifying the consumption intensity and representative patterns of the customers. This is done as follows:

- *Consumption intensity:* This clustering is applied on the representative daily load profiles of all the customers without normalization, $(\hat{X}_1, \hat{X}_2, ..., \hat{X}_n)$. The representative load profile of a customer i, $\hat{X}_i$, is estimated by the mean of the hourly consumption:

$$\hat{X}_i = \frac{1}{N} \sum_{n \in N} X(n) \tag{2}$$

  where N is the number of days in a time series, and $X(n)$ is the n-th day's daily load profile, $X(n) \in \mathbf{R}^{24}$.

- *Representative patterns:* This clustering is applied on all the normalized daily load profiles of all customers. For a given customer i and the n-th day, the normalized load profile is defined as

$$X_i^*(n) = \frac{X_i(n)}{S_i(n)} \tag{3}$$

  where $X_i^*(n)$ and $X_i(n)$ denote the 24h-period vectors of normalized and raw hourly consumption data, respectively. $S_i(n)$ is a scalar value representing the total consumption of the day n. The clustering on the normalized load profiles will only discover the patterns regardless of the consumption intensity.

3.1.1. Number of clusters K

The challenge of using the K-means algorithm is that it requires the user to provide the initial number of clusters K, and it is often difficult to determine the optimal number of clusters. There have been a number of approaches proposed, such as the random initialization for gaussian mixture models criterion in [33], using gap statistic to estimate the number of clusters [34], and an information measure for classification [35]. In this paper, the number of clusters for load intensity and consumption pattern are chosen according to the Bayesian Information Criterion (BIC) [36], which is based on the following considerations. In clustering, adding more clusters will decrease the variance (and increase the Bayesian likelihood). To avoid constantly adding centroids, the BIC works by penalizing the log-likelihood more when the complexity of the model (e.g. the number of parameters) increases [36]. Therefore, this approach can search for different values of k, score each clustering model according to the BIC value, and determine the optimal number of clusters. BIC is defined as

$$BIC(C|X) = \mathcal{L}(X|C) - \frac{p}{2} log(n) \tag{4}$$

where $\mathcal{L}(X|C)$ is the log-likelihood of the data item of X belonging to a cluster C, and $p = k(d+1)$ is the number of parameters in the C with dimensionality of d and k cluster centroids.

3.1.2. Distance metrics

The clustering is based on the distance to decide which cluster an element belongs to. The Euclidean distance is one of the most commonly used metrics by clustering algorithms, which is defined as the following.

$$d(X,Y) = \|X - Y\|_2 = \sqrt{\sum (X-Y)^2} \tag{5}$$

where X and Y are the two vectors with the same dimensionality. In the clustering of consumption intensity, they are the 24h-vectors representing the hourly consumption values of the day. This paper uses an improved distance metric derived from the Euclidean distance to discover representative patterns. It is called *KSC-distance* [37], which is suggested by [38] to discover energy consumption patterns with a minor shift on the time scale. The KSC distance is defined as

$$d(X,Y) = \min_q \|X_q - Y\|_2 \tag{6}$$

where X_q is a shifted version of X by the amount q that minimizes the instance. KSC-distance can remedy the "double-penalty" problem of the peaks that are only slightly different in timing when the peaks occur [38]. Figure 2 shows the example where there are two vectors with the peak value of $1.0kWh$. If q is shifted to the left for one hour, i.e., -1, the distance d will be 0, otherwise $\sqrt{2}$. This means that the difference between the two patterns regarding a slight shifting along the time scale can still be classified into the same cluster. This can solve minor pattern shifting problems. For example, a person getting up late with the morning peak at 8–9 am, instead of his/her usual morning peak at 7–8 am. The slight variance on the time scale should not lead to a wrong classification.

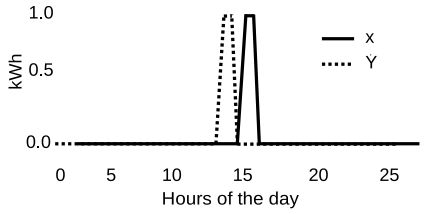

Figure 2: Double-penalty problem

In this paper, the daily pattern is shifted between -1 and 1, i.e., the last hour of the previous day and the first hour of the next day. The distance is computed iteratively for the three shifts, and the minimal one is used for the clustering.

3.2. Correlation study based on the consumption intensity

The customer groups in terms of consumption intensity are identified by the clustering analysis without normalization. A correlation study is conducted to investigate the influencing factors to each of the consumption intensity groups based on the clustering results. The influencing factors include the characteristics of buildings and occupants. The purpose of this analysis is to determine the impact of building-related parameters on the heating consumption intensity, since these parameters are commonly used in national building databases to classify building stocks. Therefore, this study is to find out whether these parameters are representative to be used in classification schemes when the focus is on energy use intensity. Furthermore, occupant-related parameters are investigated to determine if they are suitable for characterizing heating consumption classification. The reason for selecting such parameters is that occupant behavior introduces large stochasticity to residential space heating demand [39, 40]. This study is first done in an intuitive way simply using bar charts, followed by a logistic (binomial) regression analysis.

Logistic regression refers to a stochastic model in which the conditional mean of the dependent dichotomous variable (usually denoted $Y \in \{0, 1\}$) is the logistic function of an affine function of the vector of independent variables (usually denoted $\boldsymbol{x}$). That is,

$$E[Y \mid \boldsymbol{x}] = \sigma(\boldsymbol{c}^T \boldsymbol{x}) \tag{7}$$

for some unknown vector of coefficients $\boldsymbol{c}$ and where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the logistic function. The logistic regression finds the vector of coefficients $\boldsymbol{c}$ that maximizes the likelihood of the observations. The relevance of studied characteristics is quantified by studying the significance of the coefficients.

3.3. Variability study based on the representative patterns

Based on the representative patterns discovered by the clustering with normalization, a variability study is conducted to investigate the consumption behavior change over time for households. In this study, each of the patterns is represented by an alphabet, and the daily consumption patterns of each customer will, therefore, be labelled as a sequence of the alphabets. Based on the alphabet sequence, the customer behavior changes over time will be quantified as the entropy in this study. Variability reflects the consumption habits of customers, and has a determining impact on operation cost for utilities. Utilities can rely on it to design energy efficiency programs for the relevant types of the users or neighborhoods that contribute to the variability. This study also provides the information for utilities to provision the supply of energy. The concept of entropy is hereby adopted to quantify variability. Entropy originally comes from the physics field and can be generally perceived as a measure of disorder. Entropy has been widely used in computer science and information theory mainly to characterize the uncertainty in the data [41, 42]. In this case, a higher entropy value means that the user consumption is more variable, i.e., more unstable or less predictable. Otherwise, it is more predictable. There are typically two approaches for estimating the entropy: one is believing that the appearance of an alphabet α is uncorrelated, while the other is believing that the appearance is serially-correlated [38], i.e., the appearance of an alphabet for the next day depends on its previous days. For the first approach, the entropy can be calculated by

$$E^{uncor} = -\sum_{\alpha} p(\alpha) log p(\alpha) \tag{8}$$

where $p(\alpha)$ is the posterior probability of an alphabet α. The probability can be estimated by

$$p(\alpha) = \frac{\#\{S(n) = \alpha\}}{N} \tag{9}$$

where the numerator is the count of an alphabet α in the sequence, and the denominator N is the length of the sequence, i.e., the number of days. The second approach estimates the entropy using the Lempel-Ziv compression algorithm [43]. The entropy is estimated as:

$$E^{cor} = (\frac{1}{N} \sum_{n} L_n)^{-1} log(N) \tag{10}$$

where L_i is the length of the shortest sub-sequence at position n which doesn't previously appear from position 1 to $n-1$.

4. Analysis and Results

This section will first describe the data, followed by a descriptive analysis of the data. Then, this section will do the clustering-based analysis to study consumption intensity, load transition, and variability.

4.1. Data

The district heating consumption data are from 8,293 single-family households in Aarhus City, Denmark, with hourly time resolution. Single-family houses represent 44% of the total residential households in Denmark [44]. The readings were collected from different starting dates, with the earliest being March 02, 2009, but all end on November 29, 2015. These data are private and owned by the district heating supplier of the area, Aarhus Affaldvarme (AVA). AVA provided the authors with the data for research purposes after applying some anonymization techniques like removing personally identifiable information such as names, addresses and social security numbers to ensure the sources' privacy. The additional data include building information from the Danish National Building Register (BBR) such as the construction years, sites and areas. These data are publicly available at [45]. BBR is a nationwide register which includes the data of the majority of Danish buildings and households. It contains the information about 1.6 million properties, 3.8 million buildings and 2.7 million dwellings and commercial units [46]. It was originally established in 1977 by collecting the information from building owners via questionnaires. Since then, it has been updated by local authorities and by citizens [47]. Specific customer information was also provided to the authors, including family sizes, date of birth of residents, addresses and the dates of moving in and moving out.

4.1.1. Descriptive analysis of data

Initially, a descriptive analysis is conducted to visualize the time series of district heating consumption data. Figure 3 shows three typical consumption time series (high, medium and low) in the year 2014. The figure indicates that most of the heating consumption occurs in the winter period between October and April of the following year, while there is very little heating consumption during the summer period May to September. This is in line with the typical assumption for the Danish heating season, starting on October 1st and ending on April 30th [48]. Figure 4 represents the average hourly consumption of all households for weekdays and weekend days during January. The results show that the peak loads occur in the morning and evening on weekdays, while the morning peak is occurring later – approximately 3 hours – in the weekends. Based on these results, it can be observed that there are different consumption intensity levels. District heating data follow certain patterns and seasonality, which may be due to the weather conditions, building performance, living habits of customers and/or something else. In the following, this paper will conduct an exploratory analysis based on the clustering to study consumption intensity, load transition and variability.

4.2. Clustering for studying consumption intensity

In the following, an exploratory analysis will be conducted using the proposed clustering methods. Initially, the clustering is based on the daily load profiles without normalization. A daily load profile of day n is denoted as $X(n) = \langle x_0, ..., x_h, ..., x_{23} \rangle$, where x_h represents the consumption of the hour, h.

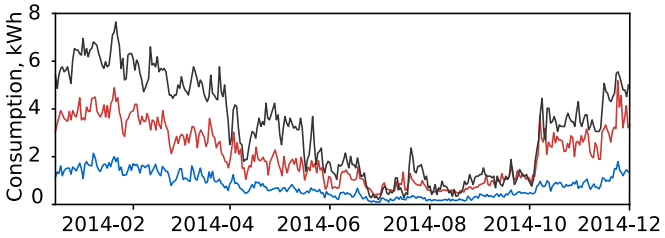

Figure 3: Three typical examples of heating consumption time series

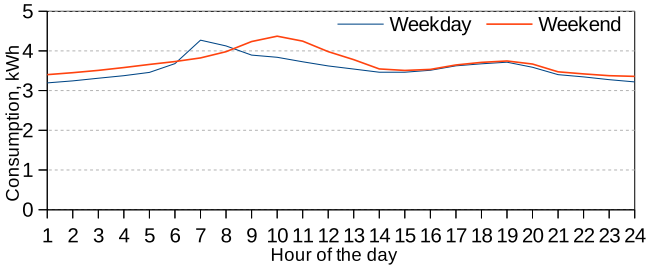

Figure 4: Average daily heating consumption patterns

According to the proposed methods, the number of clusters is defined by using the BIC approach. Figure 5 shows the evolution of BIC values with the increasing number of clusters. The BIC value decreases up to 5, $k = 5$ and increases progressively thereafter. Therefore the appropriate number of clusters is 5.

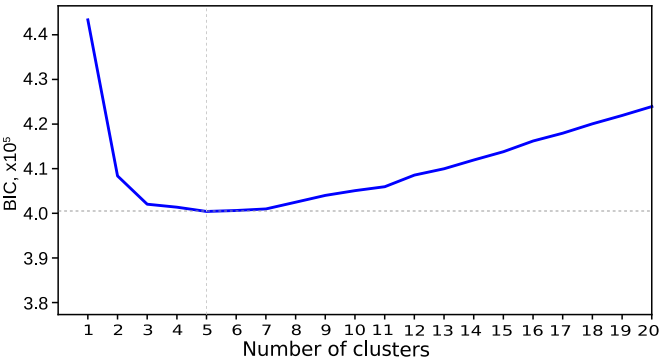

Figure 5: Determining the optimum number of clusters

The clustering, therefore, segments the customers into five consumption groups. The silhouette score [1] [49] of the clustering is 0.641, indicating that the clustering achieves a satisfactory result in terms of intra-cluster cohesion and inter-cluster separation. The clusters are ordered by their daily average heating consumption, and labelled using the alphabet from A to E, which also represents the consumption levels ranging from the lowest to the highest. The purpose of this clustering is to identify the customers with a similar *consumption pattern and consumption intensity*. This information is valuable to utilities to target the customer for offering personalized services, e.g., giving energy-saving recommendations or heat emission system

[1] Silhouette score is a measure for evaluating clustering results according to cohesion within a cluster, and the separation to other clusters. The score ranges from −1 to +1, where a high value indicates a good matching [49].

replacement. Figure 6 shows the clustering results. In each cluster, the bold red line is the cluster centroid which is the representative group load profile, while the other lines are the mean daily load profiles of the customers in the cluster. Cluster *A* represents 24.2% of the households having the lowest heating consumption among all households. The cluster load profile is rather flat, having a slight peak between 5am and 8am. Cluster *B* represents the majority of households, which is 36.2% of the total examined stock. The heating consumption is a bit higher, while the cluster load profile is rather constant with a small variation. The daily heating consumption of the centre of cluster *B* does not exceed 2.5 kWh. Cluster *C* exhibits a similar pattern as *B*, with the heating consumption being a bit higher. The morning peak occurs between 6am and 8am, while the evening peak is very weak. Cluster *D* represents 11.2% of the households, with the daily heating consumption ranging from 2.7 to 4 kWh. Cluster *E* indicates the highest heating consumption among all households, representing only 2.1% of all dwellings. The morning and evening peaks are more pronounced, with the former occurring between 6am and 8am and the latter between 5pm and 8pm.

The hourly readings of the households are now explored in each cluster using the boxplot method. The boxplot method can unveil the descriptive statistics regarding dispersion and skewness, as well as outliers of the data [50]. Figure 7 illustrates the results (the line in the box represents the median). In each cluster, there are some outliers that lay beyond the quartiles by 1.5 interquartile range. The distributions of the clusters A, B and D are more skew for their median lies at the quartile boxplot. The boxplots depict the distribution of the data in each cluster (the length of the box represents the variance, and the standard deviation range from 8.1 to 24.9 kWh). Cluster *A* has the lowest variability and consumption, while Cluster *E* has the highest variability and the highest consumption.

It is now further investigated if there exists any relevance between the energy consumption levels and building and occupant characteristics in an intuitive way. The analysis uses bar charts to determine the effect of floor area, the age of building and number of occupants on the heating load profile (see Figure 8). Visually, the results have shown that buildings with large floor areas and old buildings have high heat demand on average. Looking at Figure 8a, the average floor area is higher for cluster *E*, which represents the high heat demand profiles, as expected, since the heating consumption data are not normalised per floor area. Furthermore, it can be observed from Figure 8b that the older a building is, the higher heat demand it has. Cluster *A*, which represents the lowest heat demand, includes buildings with the lowest average age of buildings. The effect of the number of occupants is not as clear among the different clusters (Figure 8c).

The relevance is now quantified by the coefficient significance using logit regression analysis. The following features are selected: building age, floor area, the number of adults, the number of teenagers, and the number of children in a household. In this case, the independent variables are the house and occupant features. In logit regression, the dependent variable, y, only takes two possible values, 0 or 1. In this analysis, y corresponds to 1 or 0 if a sample is belonging to a cluster or not. For example, for the logit regression of cluster *A*, if a customer belongs to this cluster, then $y = 1$, otherwise, $y = 0$. The coefficients of the regression, along with the odds ratio for the coefficient, the Wald p-value and z are calculated and presented in the following. Table 1 presents the regression analysis results of the attributes of time-patterns for the five identified clusters. According to the Pseudo R^2, the models of the high energy consumption groups *D* and *E* have a better fit than the lower energy consumption groups (0.2-0.4 represents the good fit, which is equivalent to 0.7-0.9 in traditional R^2 [51]). According to the p-value, the explanatory variables, floor or building area and age of the building, are the most significant ones in all segments. The number of teenagers in the household appears to be positively associated with the low and relatively constant heat load represented by Cluster *A*. This may be attributed to teenagers spending less time at home compared with other age groups. In the high load segment, the number of teenagers changes 1.451 times the odds of a household to belong to that cluster. When looking at the medium load segment, the number of children changes 1.156 times the odds of a dwelling to belong to that cluster. The effect of the number of adults is not easy to interpret. Adult occupant behavior varies a lot and does not follow a steady pattern due to different thermal comfort preferences.

Figure 9 shows the daily load profile of all customers after quantification of the clustering differentiated between weekdays (a) and weekends/holiday (b). All five segments show a similar trend for weekdays, but not for weekends. In particular, the magnitude of heating consumption differentiates the five clusters on weekdays. All of the five clusters show a remarkable morning peak. Furthermore, the segments of high and medium load profiles are characterized by a morning peak a bit earlier in the day (around 7am), while in the two lower load profile segments the morning peak occurs approximately one hour later. In addition, cluster *E* which represents high load profiles has two soft evening peaks at 6pm and 9pm. Looking at the load profiling for weekends, the trends are less clear. Occupants do not follow a steady schedule during weekends and holiday. Cluster *D* which is the segment with the second highest heating consumption, has a striking morning peak at 9am. The behavior of the rest of the clusters is quite similar to each other and they show softer peaks during the day.

Figure 10 shows how a typical customer's daily heating consumption profile changes cluster throughout the year (2014). In the customer segmentation process in Section 3.1 this customer is in cluster *E*. It means that this customer is classified into the highest heating consumption group. However, interestingly, the customer changes consumption pattern from day to day throughout the year, which means that this customer load profile fits better with another cluster the next day. The figure shows which cluster the customer is associated with at each day within the year. Although this customer on average is a high load profile consumer, the customer is fully associated with the low load profile cluster *A* during the warmer summer months, from May to September. During autumn in October and spring in April and half of May, the customer is associated with the medium heating consumption cluster *C*. This is in the two tran-

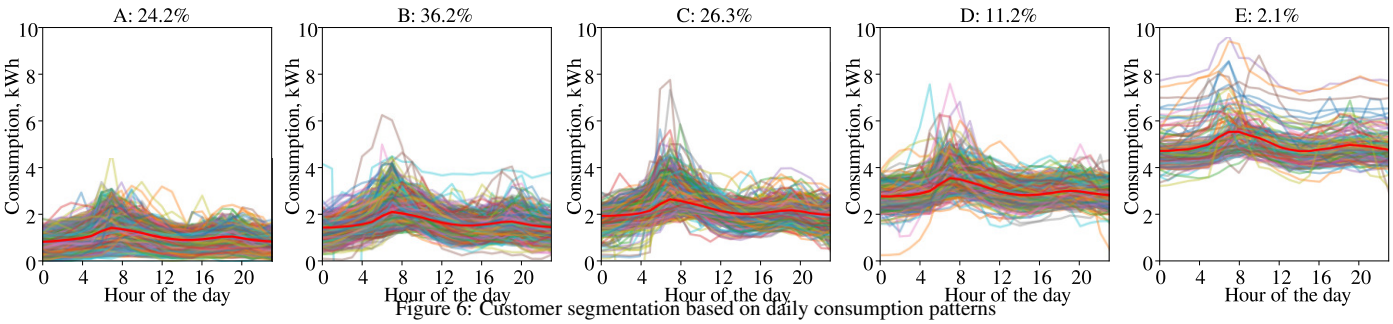

Figure 6: Customer segmentation based on daily consumption patterns

Table 1: Logit regression results for load profile segmentation

Explanatory variables	A					B				
	Coef.	Odds	std-err	z-value	p-value	Coef.	Odds	std-err	z-value	p-value
Intercept	2.476	11.900	0.6931	3.573	0.00035***	0.665	1.945	0.536	1.240	0.214
Building area	-0.010	0.989	0.00272	-3.899	0.00009***	-0.004	0.996	0.0022	-1.727	0.084.
Building age	-0.038	0.962	0.0052	-7.318	0.000***	-0.013	0.987	0.0037	-3.369	0.0007***
No. of adults	0.055	1.057	0.0786	0.707	0.479	-0.049	0.952	0.0648	-0.757	0.448
No. of teenagers	-0.549	0.577	0.2191	-2.509	0.012*	0.063	1.065	0.1424	0.443	0.657
No. of children	-0.094	0.909	0.1282	-0.736	0.461	-0.048	0.952	0.1063	-0.454	0.649
	Df=5 Log-likelihood=-90.4 Pseudo R^2=0.0526					Df=5 Log-likelihood=-140.7 Pseudo R^2=0.075				

C					D					E				
Coef.	Odds	std-err	z-value	p-value	Coef.	Odds	std-err	z-value	p-value	Coef.	Odds	std-err	z-value	p-value
-2.276	0.103	0.579	-3.930	8.45e-05***	-5.506	0.004	0.8423	-6.537	6.256e-11***	-12.919	2.449e-6	2.125	-6.078	1.211e-09***
0.002	1.002	0.00237	0.771	0.440	0.011	1.011	0.00327	3.414	0.0006***	0.036	1.037	0.007	4.724	2.307e-06***
0.016	1.016	0.00384	4.130	3.614e-05***	0.031	1.031	0.00530	5.898	3.664e-09***	0.047	1.048	0.0103	4.567	4.946e-06***
-0.066	0.936	0.0693	-0.959	0.337	0.121	1.129	0.0907	1.339	0.180	0.106	1.111	0.219	0.483	0.628
0.235	1.265	0.1465	1.605	0.108	0.000	1.000	0.2057	0.002	0.997	0.372	1.451	0.454	0.819	0.412
0.145	1.156	0.114	1.303	0.192	-0.058	0.943	0.1623	-0.361	0.717	-0.372	0.689	0.476	-0.782	0.433
Df=5 Log-likelihood=-140.8 Pseudo R^2=0.102					Df=5 Log-likelihood=-101.8 Pseudo R^2=0.214					Df=5 Log-likelihood=-19.3 Pseudo R^2=0.384				

Significance codes of p-value: 0.1', 0.05*, 0.01**, 0.001***

sition months between heating season and non-heating season where there is a medium heat demand. The customer is actually only associated with the high heating consumption cluster, *E*, during January and December, when ambient temperatures are low, probably sub-zero degrees Celsius. The customer is mostly associated with cluster *B* and *D* in February through April, as well as in November and December. The share of the time this customer is associated with the second lowest heat load cluster *B* during winter months is quite high and that indicates a well-insulated house that can possibly benefit significantly from solar gains. The relatively low heat loads can also be associated with specific calendar events, such as holidays, when occupants are not home. It should be noted that this house was built in 1929 and its floor area is 150 m^2. Therefore, it is an old house, which explains the high heating consumption during the winter months.

4.3. Clustering for studying load transition

This subsection investigates seasonal load transitions. Figure 11 presents the evolution of consumer behavior from one cluster to another over the twelve months in the year of 2014. First, this figure shows the clusters distribution of all households for each month. This can be explained by the behavior changes that are dependent on the ambient temperature and on calendar events. In particular, cluster *D*, which corresponds to the second highest heat load segment, represents the majority of the households during the winter months. It should be noted for comparison that cluster *D* represented 11.2% of the customers and cluster *E* 2.1% of the customers. The majority of the customers in January moved to clusters with lower heating consumption in February, namely cluster *B* and *C*. The proportions of the clusters, *D* and *E*, further decrease while moving towards the spring season. Cluster *C* is the dominating cluster in March, which is a transition month from the cold winter months to the non-heating season. During the summer season, the majority of the customers are represented by cluster *A*, which has the lowest heat load and a fairly constant daily pattern. Similar conclusions for the load transitions can be drawn looking at the autumn months.

Furthermore, the transition probability from one cluster to the others has been calculated to quantify the changes in heating consumption and consumer behavior (see Table 2). The probability in this table is the statistic data of the transition between daily patterns for all the customers in the year of 2014. The transition probability between months can also be calculated in a similar way. The table shows that the probability of remaining in the same state (cluster) is higher than transiting to another state. The transitions to the adjacent states are also

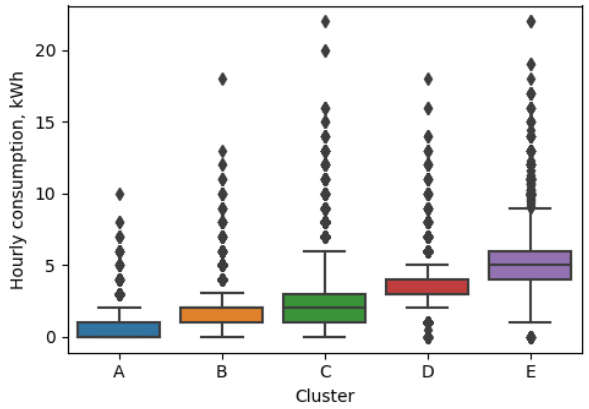

Figure 7: Boxplot of hourly consumption distribution in each cluster

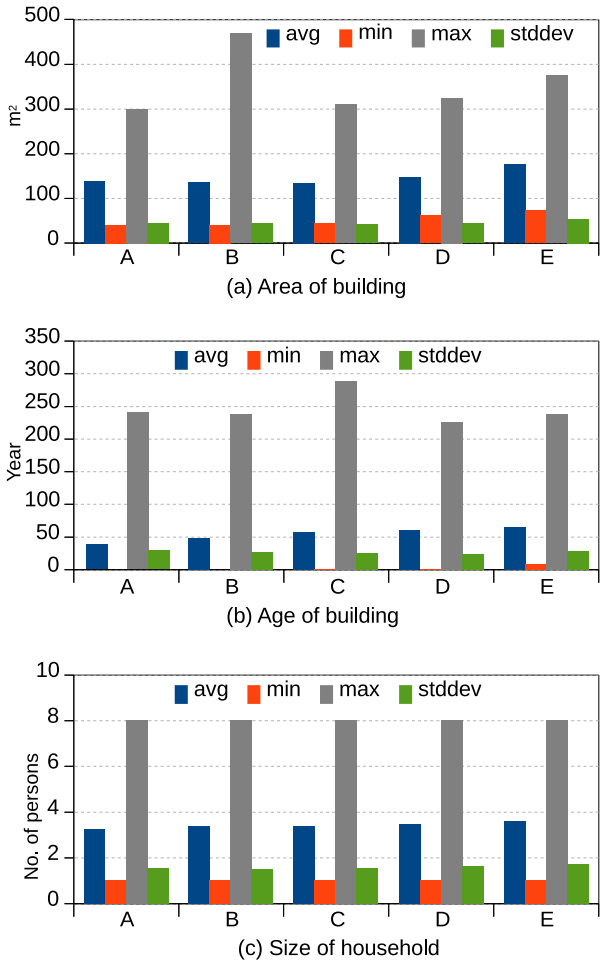

Figure 8: Statistical indicator of each group

detected, which depends on many factors, such as the changes of ambient temperature.

4.4. Clustering for studying consumption pattern variability

To study the consumption variability regarding the daily patterns, this paper first applies clustering on the normalized patterns of all days of all the customers, then quantifies the variability of each customer based on the clustering results. The normalization is conducted according to the Equation 3. In this

Table 2: Transition probability between clusters

Clusters	*A*	*B*	*C*	*D*	*E*
A	0.9246	0.0624	0.0109	0.0019	0.0003
B	0.0990	0.7839	0.0327	0.0833	0.0012
C	0.1521	0.3037	0.3568	0.1716	0.0158
D	0.0046	0.1204	0.0294	0.7976	0.0479
E	0.0026	0.0065	0.0075	0.1816	0.8018

clustering, the chosen number of clusters is nine with regards to the BIC value (see Figure 12). The silhouette score of the resulting clusters is 0.396, which is a bit lower than the clustering on un-normalized data. The reason is that the KSC distance metric is used in the clustering, and the patterns within the same cluster are more cluttered. Figure 13 shows the clusters ranked according to the percentage of the patterns in each cluster from the highest to the lowest. The clusters are labeled with the alphabets *A* –*I* (note that this labeling is different to the previous section which represents consumption intensity levels). As shown, cluster *A* represents the largest share of daily load patterns, 81.1%, where the patterns with morning and evening peaks are almost evenly distributed (as seen from the crowded lines of patterns). Since the KSC distance is used in the clustering, the minor shift of patterns are classified into the same cluster, and the centroid is flattened (represented by the bold red line). Cluster *B* has a pronounced morning peak around 9am (merged due to the use of KSC distance), and an evening lower peak around 7pm. Cluster *D* also exhibits two peaks in the afternoon and evening respectively, while the remainders have one peak, which differentiates in appearance time during the day.

Based on the identified patterns, variability is quantified by the entropy metrics. The variability can be used to provision energy supply, optimize the design and operation of the future energy system by utilities. Depending on whether the patterns are serially correlated or not, the entropy is computed by the Equations 8 and 10, respectively. The distributions of the entropy are displayed in Figure 14. The figure indicates that consumption at the individual level seems to be serially-correlated since in general $E^{cor} < E^{uncor}$, i.e., the consumption at the present depending on the past. The value of E^{cor} is much lower, which shows that the appearances of daily patterns are serially correlated. This can also be verified by the transition Table 2 which shows that the probability between two identical states is much higher. The entropy for the large portion of consumers is 0.25 for the correlated, and 1.25 for the uncorrelated.

Figure 15 compares the average entropy for the customer groups identified by clustering approach in Section 3.1. The result shows that the lower consumption intensity groups have higher variability in terms of the entropy values, while the higher consumption intensity groups have lower variability. These findings can help utilities to identify the consumption of which customer or customer group are more predictable (or stable), and to identify the customers, for example, to participate in demand-response programs.

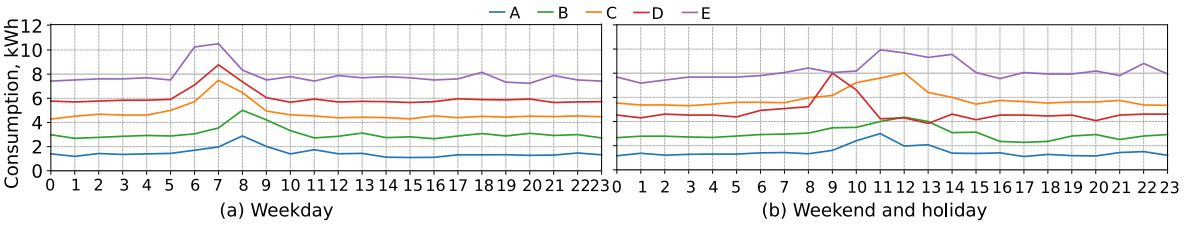

Figure 9: Cluster daily load profiles into five groups for weekday and weekend/holiday

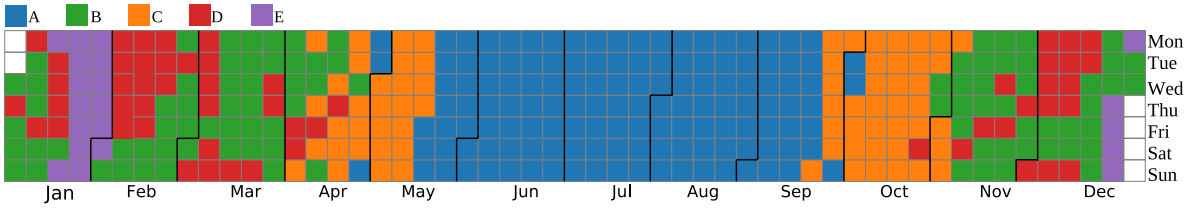

Figure 10: Load profiling of a customer over the year of 2014

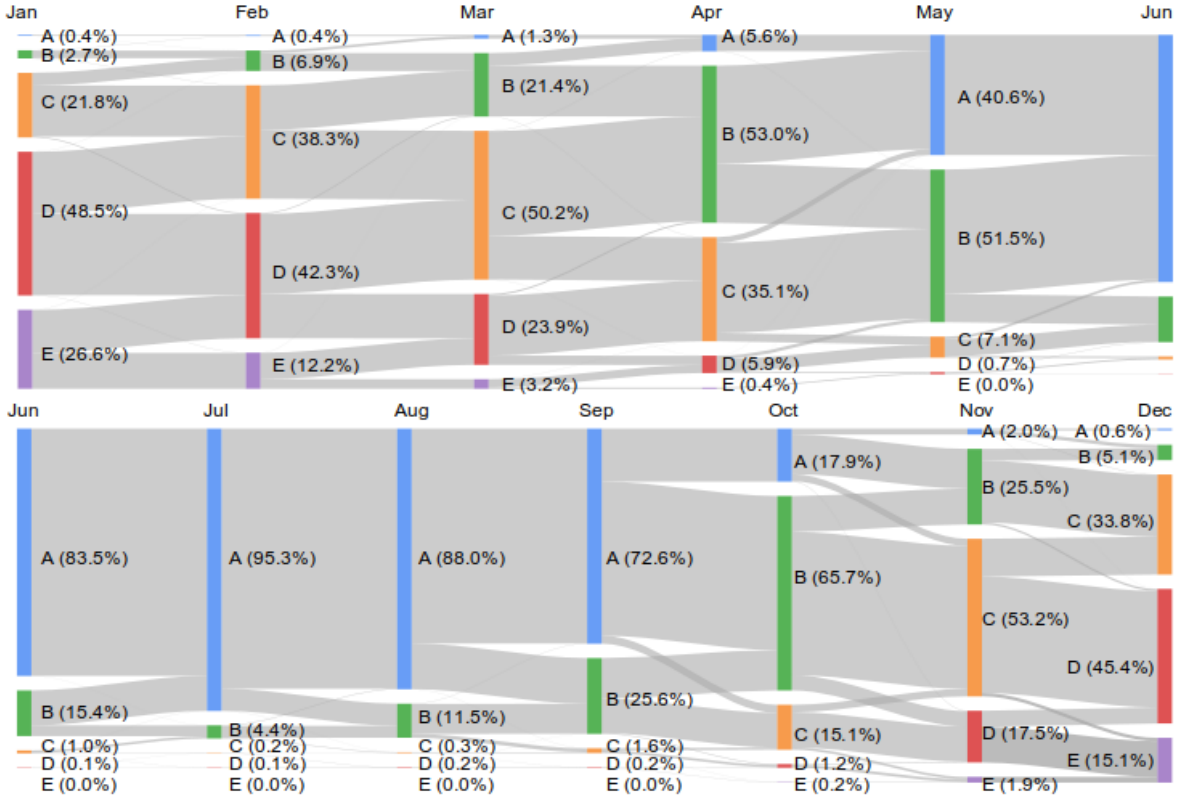

Figure 11: Load transition over the months in the year of 2014

5. Discussion

One of the main objectives of this study has been to determine and map heating consumption patterns for district heating customers in Denmark. The analysis shows that heat load profiles of Danish residential customers living in single-family houses in Aarhus can be represented by five clusters with regards to load intensity. In this case, the daily consumption patterns are fairly constant with two weak peaks, one in the morning and one in the evening. Therefore, the space heating profile shows a pattern similar to the electricity load profile that has been identified in most of the electricity consumption pattern segmentation studies, i.e., with two distinct peaks. The morning peak is more pronounced than the evening peak in the heating consumption profile, which is often the reverse for electricity. When the focus is placed on identifying specific consumption patterns, the customers are represented by nine clusters. The cluster that represents the vast majority of customers shows a fairly constant profile, while the remaining clusters are characterized by one or two peaks during the day. This finding is important as it indicates that the investigated district heating customers do not change their consumption during the day. Such information is valuable for district heating utilities to fur-

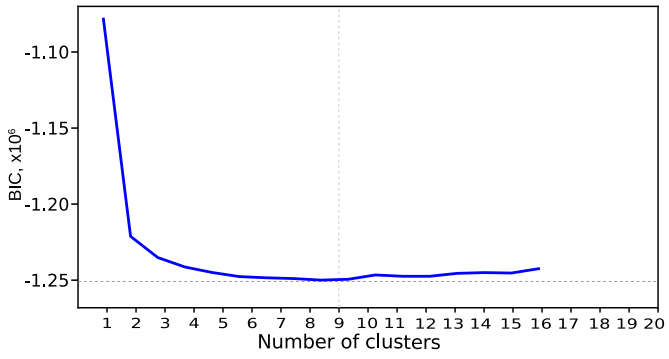

Figure 12: Determining the optimum number of clusters for normalized consumption patterns

ther optimize their network and apply advanced solutions (i.e. virtual storage on district heating network). Furthermore, the identification of the morning and/or evening peaks is important for implementing demand-side management strategies and for the possible utilization of thermal storage solutions considering the needs of the grid. During peak load periods, the grid is stressed and heating cut-off combined with activation of thermal energy storage mechanisms are solutions that could be examined in conjunction with the load profile and consumption patterns to balance the supply and demand sides.

The big data set that was used in the study increases the robustness of the method and the probability of finding statistically significant results. The sample of customers studied is very homogeneous, as the data set only includes single-family houses. Even though the results cannot be generalized for the rest of the country, it is expected that residential district heating customers in Danish urban areas show similar patterns, as the cultural habits are quite uniform in the bigger Danish cities (e.g., work schedule, desired thermal comfort conditions). If the study was to be expanded towards different building typologies, such as commercial buildings, the number of clusters to represent all customers would probably increase to accommodate more diverse heating patterns. Furthermore, the correlation of heat load patterns with building and occupant characteristics is an important step to determine if these factors can successfully reflect the heat load profiles, while partly explaining the dynamics of the space heating consumption. From a building physics perspective, more variables should be investigated, such as the time constant of the buildings, insulation level of the building envelope, infiltration rate, etc. In addition, it would be relevant to integrate a specific variable for holidays or calendar events in the model. The investigated building-related parameters (age of the building and floor area) were selected, because they have been most commonly used in national building classification schemes apart from the building types [52]. Therefore, the aim was to find out if they can accurately reflect residential heating load profiles. Results showed that they are strongly related to the load intensity of the clusters, hence they are suitable to be used for classification of Danish single-family houses (in urban areas) when the focus is placed on the energy use intensity. The family size and the age of occupants were also investigated to determine if they could substitute some of the already existing classification parameters to represent differences in occupant behavior. However, no strong correlation between the age of occupants with the consumption intensity could be drawn in the study. To draw conclusions on the effect of occupants' ages on the consumption patterns, the logit analysis can be repeated on the normalized data.

A different categorization of occupants with regards to age, such as children, adults and elderly or retired people could also be investigated, since pensioners are likely to have a different schedule than the rest of the occupants. The amount of investigated variables could also be expanded towards socio-economic characteristics, such as income per household, education level, etc., in order to identify possible targeted groups for energy efficiency programs. Data on these variables are, however, more difficult to obtain because they are not available at urban-scale building databases.

The proposed techniques can easily be implemented on large data sets. The selected algorithm, K-means has been successfully used on large data sets because of its simplicity and linear time complexity [18]. This makes it computationally attractive. This study employs in-database analysis which has a better performance than using the traditional analysis environments such as R or Matlab as the data does not have to be read out of the database. The quantitative benchmarking studies can be found at [7]. The proposed approach can be a generic solution for other energy consumption data analysis, such as electricity, water or gas. The results can help utilities for better production-side management, such as developing new pricing policies, dimensioning new parts of the network and targeting specific customers to implement demand-side management solutions. Furthermore, the results can help customers to better understand their own consumption patterns and consumption behaviors to improve their own energy efficiency. However, smart meter data are difficult to acquire at a large scale, mainly due to privacy issues. The data anonymization applied before the analysis can greatly facilitate the process of data acquisition and publishing.

The results of the current study can also be used to characterize building performance behavior at urban scale. The proposed approach and the resulting load profiles can support a scaled analysis of buildings in large urban-scale groups. The findings are also useful for whole-building and district-scale simulations to calibrate the heating consumption profiles, optimize the design of large-scale heating systems and detect anomalous behaviors.

6. Conclusions and Future Work

Smart meters have increasingly been used for monitoring heating consumption. This paper has proposed a clustering-based knowledge discovery approach for understanding residential heating consumption data, including data preparation, clustering, and data analysis. The paper applied the K-means algorithm to cluster the daily load profiles from 8,293 Danish single-family households in Aarhus. The results revealed that Danish district heating customers can be segmented into five

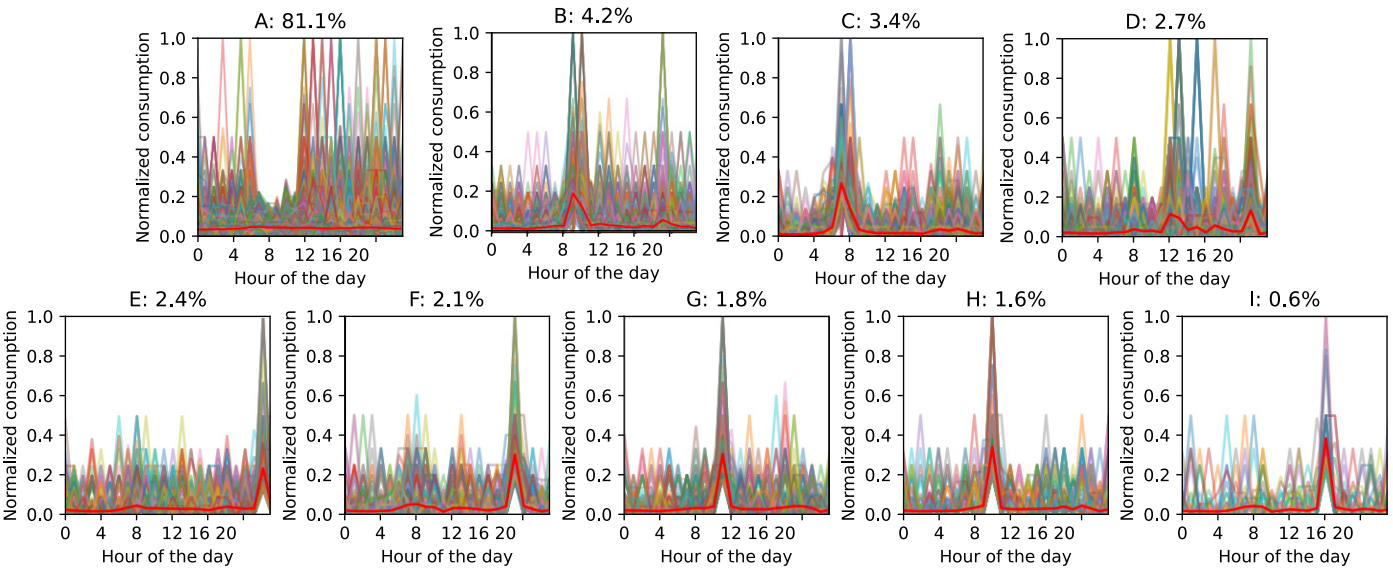

Figure 13: Clusters of normalized daily consumption patterns

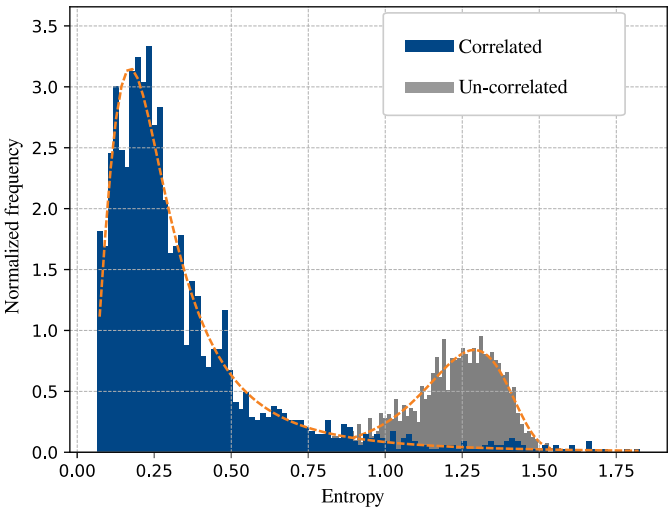

Figure 14: Distribution of entropy

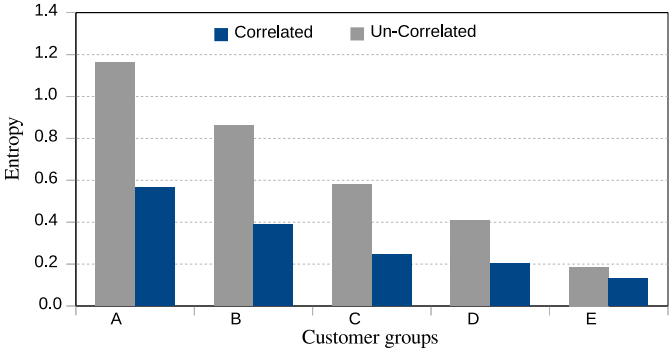

Figure 15: The average entropy of each consumption intensity group

clusters (according to the optimal number of clusters) with regards to their consumption intensity. The clusters were characterized by fairly constant load profiles with two weak peaks in the early morning and in the evening, respectively. The clusters were labeled with the alphabets, $A-E$, to represent heating consumption levels ranging from low to high. A discrimination between weekday and weekend/holiday profiles showed that weekend load profiles followed a similar pattern, but the morning peak was shifted a few hours later. To identify specific consumption patterns, a new clustering analysis was conducted on the normalized data which identified nine groups of customers, with the most dominant one showing a fairly constant profile, too. Thus, the vast majority of the examined district heating customers only made minor changes to their consumption during the day. The remaining clusters had one or two peaks in the morning or evening time, respectively. The paper also studied the correlation between heating consumption and characteristics of buildings and occupants, and used logit regression to quantify their relationship. The results indicated that building age, area and family size had a pronounced impact on the consumption, whereas the age of occupants was less pronounced. Therefore, it is appropriate to use the first two factors to categorize Danish housing stock, and particularly single-family households, in building classification schemes with regards to energy consumption intensity. In addition, the paper studied the load profile characteristics for a single customer using the clustering approach, and, based on the clustering results, illustrated the consumption transition probability over time and quantified the consumption variability using entropy methods. The results showed that the consumption patterns for each individual customer were serially correlated, and the higher consumption groups had lower variability in terms of the patterns.

For future work, clustering-based short-term energy demand forecasting will be investigated. In this study, it has been found that the majority of the customers have regular and predictable consumption behaviors. According to the transition probabilities, there is a high chance for a household to repeat the same

consumption pattern the following few days. This indicates that clustering-based load forecasting is feasible. In addition, building samples could be expanded to study the consumption patterns by including more diverse building typologies. Finally, it is also interesting to use clustering-based approach to further reveal heating consumption characteristics of buildings and occupants.

Acknowledgements. This research was supported by the Danish research project CITIES (Centre for IT-Intelligent Energy Systems in cities) no DSF1305-00027B funded by Danish Strategic Research Council. The authors would like to thank Aarhus Affaldvarme, who enabled the distribution of the smart meter data to them.

References

[1] Liu X, Golab L, Golab W, Ilyas IF. Benchmarking Smart Meter Data Analytics. Proc. of the 18th International Conference on Extending Database Technology 2015; pp. 385-396.

[2] The European Parliament and the Council of European Union. Directive 2009/72/EC concerning common rules for the internal market in electricity and repealing Directive 2003/54/EC. Official Journal of the European Union 2009.

[3] The European Parliament and the Council of European Union. Directive 2009/73/EC concerning common rules for the internal market in natural gas and repealing Directive 2003/55/EC. Official Journal of the European Union 2009.

[4] European Commission. Benchmarking smart metering deployment in the EU-27 with a focus on electricity. Report from the Commission, Brussels 2014.

[5] European Commission. Countries reports - Denmark, 2014. Available at: <https://ec.europa.eu/energy/sites/ener/files/documents/2014_countryreports_denmark.pdf> as of 2018-01-15.

[6] Darby S. Smart metering: what potential for householder engagement? Building Research & Information 2010; 38: 442–457.

[7] Liu X, Golab L, Golab W, Ilyas IF, Jin S. Smart Meter Data Analytics: Systems, Algorithms and Benchmarking. ACM Transaction of Database System (TODS) 2016; 42(1).

[8] Liu L, Nielsen PS. An ICT-Solution for Smart Meter Data Analytics. Energy 2016; 115(3):1710–1722.

[9] Kipping A, Trømborg E. Modeling and disaggregating hourly electricity consumption in Norwegian dwellings based on smart meter data. Energy and Buildings 2016; 118:350-369.

[10] Yu Z. Mining Hidden Knowledge from Measured Data for Improving Building Energy Performance. PhD Thesis, Concordia University 2012.

[11] Beckel C, Sadamori L, Staake T, Santin S. Revealing Household Characteristics from Smart Meter Data. Energy 2014; 78:397–410.

[12] International Energy Agency. Technology Roadmap. Energy-efficient Buildings: Heating and Cooling Equipment 2011; Available at <https://www.iea.org/publications/freepublications/publication/buildings_roadmap.pdf> as of 2018-01-15.

[13] International Energy Agency. Heating without global warming 2014; Available at <https://www.iea.org/publications/freepublications/ publication/FeaturedInsight_HeatingWithoutGlobalWarming_FINAL.pdf> as of 2018-01-15.

[14] Meibom P, Hilger KB, Madsen H, Vinther D. Energy Comes Together in Denmark: The Key to a Future Fossil-Free Danish Power System. IEEE Power & Energy Magazine 2013;11(5):46–55.

[15] Guha S, Rastogi R, Shim K. CURE: An Efficient Clustering Algorithm for Large Databases. ACM Sigmod Record 1998; 27(2):73–84.

[16] Theodoridis S, Koutroubas K. Pattern Recognition. Academic Press 1999.

[17] Fayyad MU, Piatesky-Shapiro G, Smuth P, Uthurusamy R. Advances in Knowledge Discovery and Data Mining (Vol 21) 1996; AAAI press.

[18] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. ACM Computing Surveys 1999; 31(3):264–323.

[19] Beckel C, Sadamori L, Santini S. Towards Automatic Classification of Private Households Using Electricity Consumption Data. In Proceedings of BuildSys '12 of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings; Toronto, Canada; 169-176.

[20] McLoughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. Applied Energy 2015; 141:190–199.

[21] Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 2012; 42:68–80.

[22] Andersen FM, Larsen HV, Boomsma TK. Long-term forecasting of hourly electricity load: Identification of consumption profiles and segmentation of customers. Energy Conversion and Management 2013; 68:244–252.

[23] Hsu D. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. Applied Energy 2015; 160:153–163.

[24] Granell R, Axon CJ, Wallom DCH. Clustering disaggregated load profiles using a Dirichlet process mixture model. Energy Conversion and Management 2015; 92:507–516.

[25] Madeira C, do Carmo R, Christensen CH. Cluster analysis of residential heat load profiles and the role of technical and household characteristics. Energy and Buildings 2016; 125:171–180.

[26] Ma Z, Yan R, Nord N. A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings. Energy 2017; 134:90–102.

[27] Mendaza IDC, Pigazo A, Bak-Jense B, Chen Z. Generation of Domestic Hot Water, Space Heating and Driving Pattern Profiles for Integration Analysis of Active Loads in Low Voltage Grids. In Proc. of Innovative Smart Grid Technologies Europe (ISGT Europe) 2013.

[28] Teeraratkul T, O'neill D, Lall S. Shape-Based Approach to Household Load Curve Clustering and Prediction. IEEE Transactions on Smart Grid 2017.

[29] Paparrizos J and Gravano L. k-Shape: Efficient and Accurate Clustering of Time Series. ACM SIGMOD Record 2016; 45:69–76.

[30] Apache MADlib: Big Data Machine Learning in SQL. Available at <http://madlib.apache.org> as of 2018-01-15.

[31] PostgreSQL. <https://www.postgresql.org> as of 2018-01-15.

[32] Azimi R, Ghayekhloo M, Ghofrani M. A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting. Energy Conversion and Management 2016; 118:331–344.

[33] Figueiredo M, Jain AK. Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Machine Intell 2002; 24(3):381–396.

[34] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J. Roy. Statist. Soc. B 2001; 411–423.

[35] Wallace CS, Boulton DM. An information measure for classification. Comput. J. 1968; 11:185–195.

[36] Robert EK, Larry W. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association 1995; 90(431):928–934.

[37] Yang J, Leskovec J. Patterns of temporal variation in online media. : Proc. of the 4th ACM international conference on Web search and data mining 2011; 177-186.

[38] Albert A, Gebru T, Ku J, Kwac J, Leskovec J, Rajagopal R. Drivers of variability in energy consumption. In ECML-PKDD DARE Workshop on Energy Analytics 2013.

[39] Sun K, Hong T. A framework for quantifying the impact of occupant behavior on energy savings of energy conservation measures. Energy and Buildings 2017; 146:383–396.

[40] O'Brien W, Gaetani I, Carlucci S, Hoes PJ, Hensen JLM. On occupant-centric building performance metrics. Building and Environment 2017; 122:373–385.

[41] Zhu C, Wang Z. Entropy-based matrix learning machine for imbalanced data sets. Pattern Recognition Letters 2017; 88: 72–80.

[42] Kiluk S. Diagnostic information system dynamics in the evaluation of machine learning algorithms for the supervision of energy efficiency of district heating-supplied buildings. Energy Conversion and Management 2017; 150: 904–913.

[43] Schurmann T, Grassberger P. Entropy estimation of symbol sequences. Chaos 1996; 6(3):414–427.

[44] Danmarks Statistik. Statistics Denmark. Available at

<https://www.statistikbanken.dk/statbank5a/default.asp?w=2021>
as of 2018-01-15.

[45] Bygnings- og Boligregistret (BBR). Danish Building Register. Available at < https://ois.dk> as of 2018-01-15.

[46] Wittchen KB, Kragh J. Danish building typologies. Participation in the TABULA project, Hrsholm, Denmark: SBi, Danish Building Research Institute, Aalborg University 2012.

[47] The Danish Ministry of Housing, 2006. Available at <http://boligejer.dk/> as of 2018-01-15.

[48] Stampe OB. Varme- og Klimateknik, Grundbog. Danvak 1992. Editors: Hansen HE, Kjerulf-Jensen P.

[49] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 1987, 20:53–65.

[50] Frigge M, Hoaglin DC, Iglewicz B. Some implementations of the boxplot. The American Statistician 1989; 43(1):50–54.

[51] Louviere, JJ, Hensher DA, Swait JD. Stated choice methods: analysis and applications. Cambridge university press 2000.

[52] Wittchen KB, Mortensen L, Hols SB, Bjrk NF, Vares S, Malmqvist T. Building typologies in the Nordic countries, Identification of potential energy saving measures. Danish Building Research Institute, Aalborg University 2012; SBi 2012:04.