Journal of Universal Computer Science, vol. 24, no. 9 (2018), 1306-1329
submitted: 30/12/17, accepted: 30/6/18, appeared: 28/9/18 © J.UCS

The Generation of Electricity Load Profiles Using K–Means Clustering Algorithm

Rūta Užupytė
(Baltic Institute of Advanced Technology
Vilnius University, Vilnius, Lithuania
ruta.uzupyte@bpti.lt)

Tomas Babarskis
(Binar Solutions, UAB, Vilnius, Lithuania
t.babarskis@binar.lt)

Tomas Krilavičius
(Baltic Institute of Advanced Technology
Vytautas Magnus University, Vilnius, Lithuania
t.krilavicius@bpti.lt)

Abstract: Accurate information about the actual behavior of electricity users is essential to the electricity suppliers in order to ensure efficient decisions in planning pricing, e.g., designing tariffs and load planning. Load profiles of customers is a straightforward source for such data, however it should be analyzed to extract relevant information. Most of the existing techniques are tested with small data sets or over short periods, which does not allow to investigate seasonality influence. We present a new methodology for the grouping of electricity customers based on the similarities of their (hourly) consumption patterns. Approach is based on the periodicity analysis and well-known clustering technique – K-means, which is applied for identification for separate users load profiles and clustering of load profiles. Values of model parameter are selected using adequacy measures. Finally, the results obtained by this methodology with a data set of 3753 electricity customers are presented, and future plans discussed.

Keywords: electricity patterns, load profiling, time–series clustering, clustering technique
Category: I.2.1

1 Introduction

Constantly growing efficiency, reliability and sustainability requirements for the electrical grid advances development of a smart grid. Roll-out of such technology allows recording electricity consumption at hourly rate and sending it to a central system at least daily. Such data could be used to improve different electricity grid parameters by exploiting knowledge about the user's behavior, e.g. profiles of users. Identification of customers groups exhibiting similar consumption patterns would allow electricity providers to design specific tariff options for the different classes of electricity customers as well as to develop a better marketing and trading strategies. From the customers point of view, benefit can be obtained by planning their electricity consumption, for example shifting it to less expensive times in order to lower their electricity bill. However, customer segmentation without any prior knowledge (about number of groups or customer's energy consumption habits) is not an easy task, new techniques are necessary in order to deal with such amount and type of data.

Different types of clustering techniques have been already used for electricity customers classification. Methodology based on the self-organizing maps (SOM) and clustering methods (K-means and hierarchical clustering) was applied on a dataset consisting of hourly measured electricity use data for 3989 electricity customers [Rasanen, 10], where only to 5% of the randomly chosen (using uniform distributions) initial time series observations were clustered. SOM was used to identify vectors that represent users consumption habits and K-means and hierarchical clustering methods were used to cluster these vectors. Results show that K-means algorithm outperforms hierarchical clustering method.

Self-organizing maps (SOM) and K-means were used in [Aickelin, 11] as well. Here, an initial data set was stratified by splitting it to weekends and weekdays (workdays), and further stratified into winter and summer seasons. Detailed analysis was applied only to one stratification, i.e. winter weekends. Each user was represented by the average load profile calculated as the mean value for each hourly reading across all readings. The results showed that K-means approach was the best in comparison with SOM and two stage process (first applying SOM and then K-means method) approaches.

Similar data stratification was applied for analyzing 15 min. resolution smart meter data for ~200 electricity users [Flath, 12]. In this case clustering was performed using K-means algorithm. Similar research [Wong, 12] based on K-means algorithm was performed using 15 - minute interval data set for 8337 households collected during two months of summer.

Research [Liu, 15] presents adaptation of K-means clustering algorithm to analyze similar behavior between customer of electricity. In this case one day's power loads with 144 observations were analyzed. Principal component analysis was used in order to get the clustering result visible. Obtained results are promising, showing the rationality and correctness of the clustering.

Paper [Poggi, 15] describes a new methodology based on high-dimensional regression models. Research data set consists of 4 225 individual customers meters, each with 48 half-hourly meter reads per day over 1 year: from 1st January 2010 up to 31st December 2010. Results revealed the necessity of separate analysis of customer load profiles for summer/winter seasons and working/non-working days.

In the mentioned studies, the customer classification was typically performed using datasets consisting of small number of electricity users (e.g., up to 2000 in [Flath, 12] or short periods of time (e.g., two summer months only in [Wong, 12])) or observations from the short time periods (up to one year). We propose a methodology to handle large electricity load time series and refine such raw data into more valuable information, i.e. user profiles. Furthermore, we propose technique to combine weekdays and weekends profiling results. Methodology presented in this paper helps to analyze electricity consumption data, though it does not provide fully automated procedures. Final decisions regarding number of profiles, minimum size of profile can be also influence by the aim of analysis or other external factors and should be confirmed by the marketing experts.

The rest of the paper is organized as follows. Section 2 presents the clustering techniques used. Section 3 describes the proposed profiling and Section 4 presents some experimental results. Finally, conclusions are given in Section 5.

2 Methodology

The proposed profiling approach can be divided into several major steps. First, we need to identify the unit of analysis (i.e. a time period over which load profile will be investigated) we want to focus our attention on. We do this by performing periodicity analysis using Lomb-Scargle periodograms. In the next phase we are trying to recognize each customer's habitual electricity consumption behavior. To detect these typical load profiles we use K-means clustering algorithm. Once we have identified these individual usage patterns, the next phase is to aggregate similar load patterns into clusters of similar consumption behavior. This is done by using K-means clustering algorithm. In order to estimate the appropriate number of clusters we use clusters adequacy indexes. These steps will be discussed in greater detail later in this paper (sec. 3). In the following subsections we provide theoretical background of methods used to analyze the data.

2.1 Lomb-Scargle periodogram

Lomb–Scargle periodogram [Scargle, 82] can be used to detect periodic patterns in unevenly spaced time series. Suppose that $\{y(t_i)\}$ is time series of n data points collected at times t_i where $i = 1, \dots, n$. In this case the observation moments are unevenly spaced, so the intervals $\delta_i = [y(t_{i-1}), y(t_i)], i = 2, \dots, n$ may have different length. Lomb – Scargle normalized periodogram can be computed from Eq. (1).

$$\hat{y}(w) = \frac{1}{2\sigma^2} \left[\frac{(\sum_{i=1}^{n} (y(t_i) - \bar{y}) \cos w\, (t_i - \tau))^2}{\sum_{i=1}^{n} \cos^2 w(t_i - \tau)} + \frac{(\sum_{i=1}^{n} (y(t_i) - \bar{y}) \sin w\, (t_i - \tau))^2}{\sum_{i=1}^{n} \sin^2 w(t_i - \tau)} \right], \tag{1}$$

where

1. $\bar{y} = 1/n \sum_{i=1}^{n} y(t_i)$ is *time series mean*;
2. $\sigma^2 = 1/(n-1) \sum_{i=1}^{n} (y(t_i) - \bar{y})^2$ is *time series variance*;
3. τ is *time delay* defined by the equation (2):

$$\tau = \frac{1}{2w} arctan \left[\frac{\sum_{i=1}^{n} sin\, 2wt_i}{\sum_{i=1}^{n} cos\, 2wt_i} \right]. \tag{2}$$

Using Eq. 1 it is easy to find a normalized power as a function of angular frequency $w = 2\pi/P$ for all the periods P. One of the simplest ways to identify the dominant time series periods is drawing the periodogram. It is easy to identify visually frequencies corresponding to the dominant spikes of the periodogram.

2.2 Data normalization

Let X be a numeric attribute with n observed values $x_i \geq 0, i = 1, \dots, n$. A value x_i of X is normalized to $\tilde{x}$ by computing

$$\tilde{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{3}$$

where $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of an attribute X. This transformation ensures that all values fall within the range [0,1].

2.3 K-means clustering

Suppose we observe $X_1, \dots, X_n \in \mathcal{R}^m$. The idea of *K-means* is to partition the $\{X_i\}$ into k $(k < n)$ clusters so that the objects within a cluster are more similar to each other than the objects in different clusters. Objects similarity can be measured using *Euclidean, Manhattan* [Han, 12] or other measures of similarity.

Stepwise K-means clustering [Han, 12] algorithm can be defined as follows.

1. Randomly select k initial objects as centroids.
2. Calculate the similarity of all objects from those k centroids.
3. Assign each object to the closest cluster.
4. Recompute each cluster centroid as the average of the objects assigned to them.
5. Repeat steps 2 – 4 until the same points are assigned to each cluster in consecutive rounds.

The objective of K-means algorithm is to minimize the squared error function:

$$E = \sum_{i=1}^{k} \sum_{X \in C_i} |X - c_i|^2, \tag{4}$$

where X – is a point in space representing a given object, c_i is the mean value of cluster C_i.

2.4 Clusters adequacy indexes

K-means clustering method requires *a priori* specification of the number of clusters (k). We apply *cluster validity indexes* to select the appropriate value of the number of clusters. The definitions of selected adequacy measures are based on the following preliminaries:

Suppose

1. $X = (x_1, \dots, x_m), Y = (y_1, \dots, y_m)$ – points in space representing given objects
2. C_i – the ith cluster
3. c_i – the centroid of cluster C_i
4. n_i – number of objects in ith cluster
5. n – the number of objects in the data set
6. k – the number of clusters

The similarity between two objects (X and Y) is defined as Euclidean distance:

$$d(X,Y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}. \tag{5}$$

2.4.1 Dunn index (DI)

Dunn index [Babos, 02] assumes that well separated clusters have a high intra-cluster similarity (between objects in the same cluster) and low inter-cluster similarity (between different clusters).

$$DI = \frac{d_{min}}{d_{max}}, \tag{6}$$

where

$$d_{min} = \min_{\substack{1 \le i,j \le k \\ i \ne j}} d_{ij}, \tag{7}$$

$$d_{ij} = \min_{X \in C_i, Y \in C_j} d(X,Y), \tag{8}$$

$$d_{max} = \max_{1 \le i \le k} D_i, \tag{9}$$

$$D_i = \max_{X,Y \in C_i} d(X,Y), \tag{10}$$

and where d_{ij} is the smallest value of similarity between two objects from different clusters, D_i is the highest value of similarity between two objects from the same cluster. The higher the value of *DI* the better is clustering.

2.4.2 Davies – Bouldin index (DBI)

DBI [Aickelin, 12] is defined as the ratio between the within cluster scatter and the between cluster separation

$$DBI = \frac{1}{k} \sum_{j=1}^{k} \max_{\substack{1 \le i \le k \\ i \ne j}} \frac{S_i + S_j}{M_{ij}}. \tag{11}$$

Where S_i is a measure of scatter within the ith cluster defined as

$$S_i = \frac{1}{n_i} \sum_{X \in C_i} d(X, c_i), \tag{12}$$

$M_{i,j}$ is a measure of separation between ith and jth clusters:

$$M_{i,j} = d(c_i, c_j). \tag{13}$$

Lower value of *DBI* indicates the better clustering. *This index has limitation: it cannot be applied when a clustering algorithm produces clusters containing single object.*

2.4.3 Mean index adequacy (MIA)

MIA [Aickelin, 12] is a *cluster compactness measure*

$$MIA = \sqrt{\frac{1}{k}\sum_{i=1}^{k} S_i}. \tag{14}$$

The smaller value of *MIA* indicates more compact clusters.

2.4.4 Cluster dispersion indicator (CDI)

CDI [Aickelin, 12] depends not only on the similarity between the objects from the same cluster but also on the similarity between all clusters centroids:

$$CDI = \sqrt{\frac{\frac{1}{k}\sum_{i=1}^{k} \frac{1}{2n_i}\sum_{X,Y\in C_i} d(X,Y)}{\frac{1}{2k}\sum_{1i,jk} M_{i,j}}}. \tag{15}$$

This measure evaluates not only the compactness of the clusters but the difference between clusters. The smaller the value of the *CDI* the better is clustering.

3 Profiling approach

The proposed profiling approach includes the following steps:

1. **Periodicity analysis.** In order to identify the unit of analysis we use the Lomb–Scargle periodogram.
2. **Data preprocessing.** Cleaning the data with missing values and applying normalization procedure.
3. **Identification of load profiles.** K–means clustering is used to find the main consumption patterns for each user during weekends/holidays and working days. The appropriate number of patterns is selected based on the results of experiments with the different number of clusters.
4. **Clustering of load profiles.** K–means are used for load profiles clustering in this step. Working days and weekends/holidays patterns are analyzed separately. In order to find the appropriate number of clusters experiments with a different number of clusters are performed and evaluated using different clustering quality measures.
5. **Results consolidation.** User groups are formed by combining weekends/holidays and working days patterns and choosing the most commonly observed combinations.

The stages of the process are explained in detail below.

3.1 Periodicity analysis

The first step is identification of analysis unit. We use Lomb-Scargle periodogram

[Scargle, 82] to identify periodicity because a part of the data is unequally sampled, i.e. some readings are missing due to the communication errors. A spectral analysis is performed for each electricity consumption curve in order to highlight the most intense periods. The most commonly observed period is used as an analysis unit, i.e. electricity load profiles are generated over this period.

3.2 Data preprocessing

The data preprocessing phase includes several phases:

1. *Cleaning process*: time series with missing values are removed[1].
2. *Data normalization*: min – max normalization (subsection 2.2) is applied, using the results of periodicity analysis, i.e. data is normalized within the most common period readings. This step ensures that clustering will be performed based on the shape of the pattern and not on the total usage. Domain experts suggest that it is more important to analyze similar load patterns according to a shape-based criterion. In this case clusters can be interpreted in terms of timing of higher and lower discretionary demand.
3. *Data partitioning*: based on the idea that users may have different energy consumption patterns during weekends and other public holidays we decided to divide data set into two smaller sets: weekends/holidays and working days observations.

3.3 Identification of customers load profiles

The goal of this step is to identify *typical patterns representing each customer's electricity demand during weekends/holidays and working days.*

Each user is represented by the data set $\mathbf{X} = \{\mathbf{X}_j^{i,d}\}$ where i denotes specific user and $d \in \{w, n\}$ is an index with two values: w stands for working days observations and n for non-working. $X_j^{i,d} = \langle x_1, x_2, \dots, x_m \rangle$ is a vector representing hourly measured electricity usage observations during the identified period (subsection 3.1). E.g., if we are looking for typical twenty - four hours patterns then $m = 24$ and $X_j^{i,d}$ is 24 values vector depicting hourly measured electricity consumption on specific date j.

Each data set ($\mathbf{X}^{i,w}$ and $\mathbf{X}^{i,n}$) is separately analyzed by applying K–means clustering algorithm (subsection 2.3). In order to identify the most appropriate number of clusters we suggest performing several experiments with a different number of clusters $k = 2,3,4,5$. Some of industrial companies may have different patterns during the week. The maximum number of 5 clusters was selected in case there would be 5 different patterns – each for every working day. The following profiling approach is based on the idea that each user has one consumption pattern during weekends/holidays and one during working days. We plan to examine more complicated cases, when a user might have more than one consumption pattern, in future. While, in this paper users' profiles for working and non - working days are

[1] In the future we are planning to extend the proposed algorithm for dealing with incomplete data. In order to calculate the distance between data that contain missing values we are going to use partial distance strategy.

identified as the centroids of the larger cluster when $k = 2$.

3.4 Classification of load profiles

In this stage we have two data sets $\mathbf{X}^w$ and $\mathbf{X}^n$ representing customers load profiles for working and non-working days. Vector $X_i^d = \langle x_1, x_2, \dots, x_m \rangle$, $d = w, n$ defines ith user's load profile for working days or weekends/holidays (based on the value of d). The value of m can be chosen according to the desired period of load pattern representation, typically, by using 24 hours period ($m = 24$), one-week period ($m = 24 \cdot 7$) and etc.

Each data set is analyzed separately by applying K–means clustering algorithm with different number of clusters $k = 2, \dots, l$. Value of l can be identified intuitively or based for example on the rule of thumb [Bibby, 79]:

$$l \approx \sqrt{n/2}, \qquad (16)$$

where n is number of objects.

In order to compare clustering results obtained using different number of clusters we propose to use several clustering adequacy measures: *MIA*, *CDI*, Davies–Bouldin and Dunn indexes (subsection 2.4). Each adequacy measure is analyzed separately. It is expected that different indexes should suggest the same or at least similar number of clusters.

3.5 Merging results

The last stage in customers profiling is results consolidation. In order to combine working days and weekends/holidays results we analyze all possible combinations of working days and weekends patterns. We use a two-way frequency table to count the number of users being in the corresponding clusters. In this way we can obtain information about the most common combinations and very rare cases. Domain experts suggest that combination of load profiles can be identified as a rare if it is ten times smaller than the largest combination. Based on the specific profiling task rare combinations can be analyzed as special ones or be assigned to the larger combination. The assignation can be performed using one of the following approaches:

1. All rare combinations are assigned to the largest (the most common) profile.
2. A new profile (consisting of all rare combinations) is formed.
3. Rare combination is assigned to one of the previously defined profiles based on the similarity measure.

For example, suppose that $g_s = \{Z_{s_1}^w, Z_{s_2}^n\}, s = 1, \dots, k$ are defined profiles (k most common combinations), where s – profile number, $Z_{s_1}^w$ – the s_1th usage pattern for working days and $Z_{s_2}^n$ – the s_2th usage pattern for weekends/holidays. A rare combination $f_{ij} = \{Y_i^w, Y_j^n\}$ is assigned to the tth profile g_t, where t is defined as

$$t =_{1 \le s \le k} d(f_{ij}, g_s), \qquad (17)$$

where

$$d(f_{ij}, g_s) = \frac{m_1 d(Y_i^w, Z_{s_1}^w) + m_2 d(Y_j^w, Z_{s_2}^w)}{m_1 + m_2}. \tag{18}$$

Values of m_1 and m_2 can be identified based on one of the following assumptions:

1. working days patterns do not have influence on profile selection: $m_1 = 0, m_2 = 1$
2. non-working days patterns do not have influence on profile selection: $m_1 = 1, m_2 = 0$
3. working and non - working days patterns have the same influence on profile selection: $m_1 = 1, m_2 = 1$
4. influence of working and non - working days patterns can be expressed as ratio $m_1 : m_2$.

The selection of an appropriate approach is a marketing problem and in this paper it will not be further analyzed.

4 Experimental Evaluation

Proposed clustering approach was applied for a real-world data set consisting of 3 years (January 2011 to December 2013) hourly measurements for 3753 industrial users. Previously presented algorithm was used to partition the data set into homogeneous clusters in order to identify the groups of electricity customers that have similar patterns of electricity consumption.

All experiments were performed using R software environment for statistical computing (www.r-project.org).

4.1 Periodicity analysis

The first step is identification of analysis unit. We apply Lomb–Scargle periodogram [Scargle, 82] to identify periodicity. We perform a spectral analysis of each electricity consumption curve, see Figs. 1 – 4 for some examples. It is easy to see that the most intense peaks for different series are different, e.g. 24 hours, 168 hours (7 days) or 8768 hours (365 days). Examples provided in this paper represent the variety of frequencies observed in the analyzed time series. More examples can be found at [Examples]. Periodicity analysis revealed that 24 hours period is common for all time series, but the intensity of this period may be different. Moreover, 24 hours period is the most relevant from the perspective of marketing decisions. However, the proposed methodology is not tied to any specific period and can be used to analyze electricity consumption over different time periods. However, local maximum values of the periodograms are the same for all the time series. This leads to the conclusion that there are several periods common for all the time series, but the intensity of these periods is different. Further analysis shows that 24 hours period is the strongest or one of the strongest periods for all the time series and as a result we choose it as a unit of the analysis. In the future we are planning to perform more detailed analysis using different units of analysis, for example 12 hours, 168 hours (7 days) or 744 hours (1 month) periods.

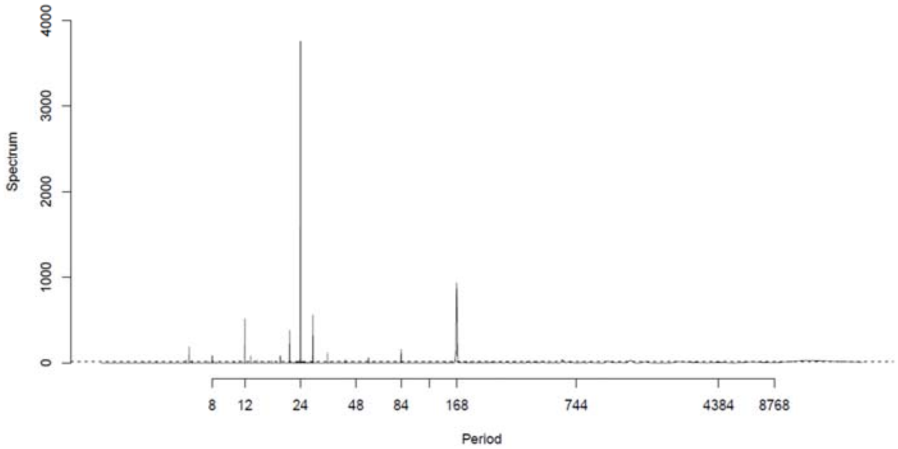

Figure 1: Periodogram with the dominant period of 24 h (selected user No. 1)

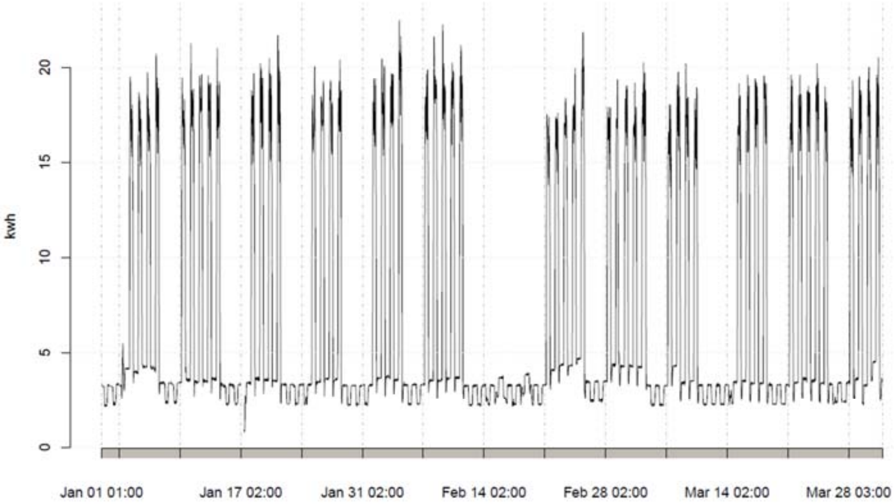

Figure 2: Electricity usage curve (selected user No. 1)

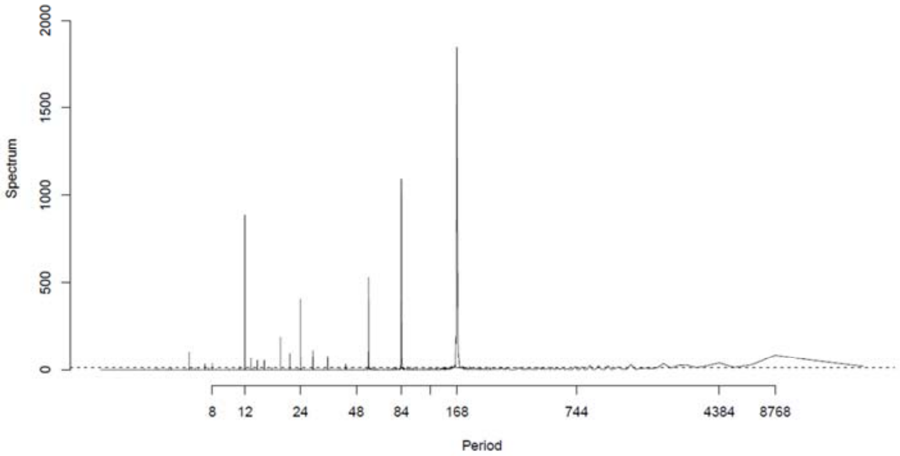

Figure 3: Periodogram with the dominant period of 7 days (selected user No. 2)

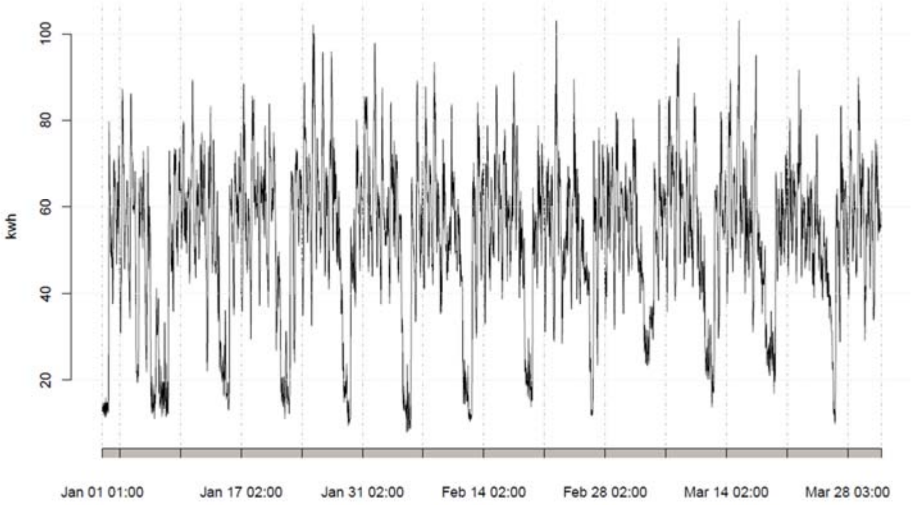

Figure 4: Electricity usage curve (selected user No. 2)

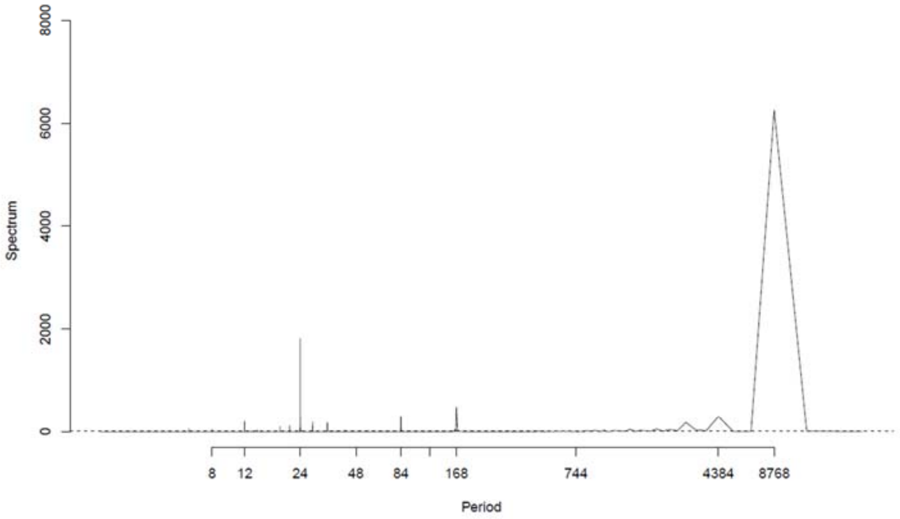

Figure 5: Periodogram with the dominant period of 365 days (selected user No. 3)

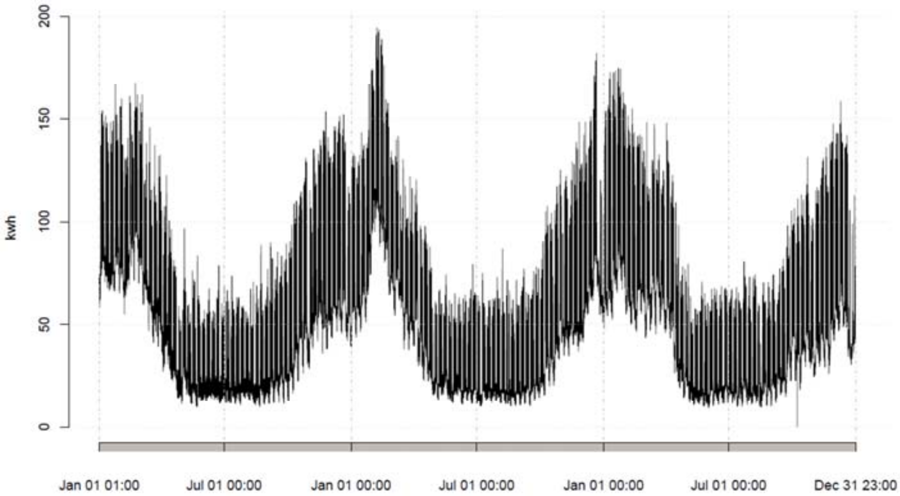

Figure 6: Electricity usage curve (selected user No. 3)

Before further analysis we perform several data preparation steps, defined in subsection 3.2. Time series with missing values are eliminated (final data set consist of 1500 user’s data) and the rest of the data is normalized within the 24 hours readings.

4.2 Identification of customers load profiles

The goal of this step is to identify *typical twenty-four hours patterns representing each customer's electricity demand during weekends/holidays and working days.* We follow the procedure from subsection 3.3.

Part of the results of experiments with a different number of clusters are provided in Fig. 7 – 12. From these graphs we can see that despite the change in the number of clusters centroids remain qualitatively unchanged. We can also observe the shift in time axis or quantitative change in consumption. E.g., in Fig. 7 the case with two clusters ($k = 2$) depicts electricity consumption during two different periods: black curve represents usage pattern during the period from April to October and black – during the period from November to March. This one-hour shift is probably caused by the daylight saving time clock shifts[2]. Another example (see Fig. 8) shows two different consumption patterns during non-working days: in case when $k = 2$ black curve depicts usage pattern on Sundays and holidays and red curve on Saturdays. These differences between consumption patterns may occur due to the shorter working hours on Sundays and holidays. Graphical example provided in Fig. 11 depicts fairly similar results: in case when $k = 2$ black curve represents consumption pattern during period from November to March and red curve – from April to October. However, in this case besides the shift in time axis we can also observe quantitative changes in consumption patterns. These differences can be influenced by change in volume of production during summer/winter seasons. A more detailed analysis is necessary for a deeper understanding of the causes of these differences. However, these observations lead to the conclusion that each user have one consumption pattern during weekends/holidays and one during working days which are identified as the centroids of the larger clusters when $k = 2$[3].

[2] Daylight saving time is the practice of advancing clocks during summer months by one hour in order to make the most efficient use of seasonal daylight. In Europe clocks are adjusted forward one hour on the last Sunday in March and adjust them backward on the last Sunday in October.

[3] With the selected data set, i.e. it could differ with other data sets.

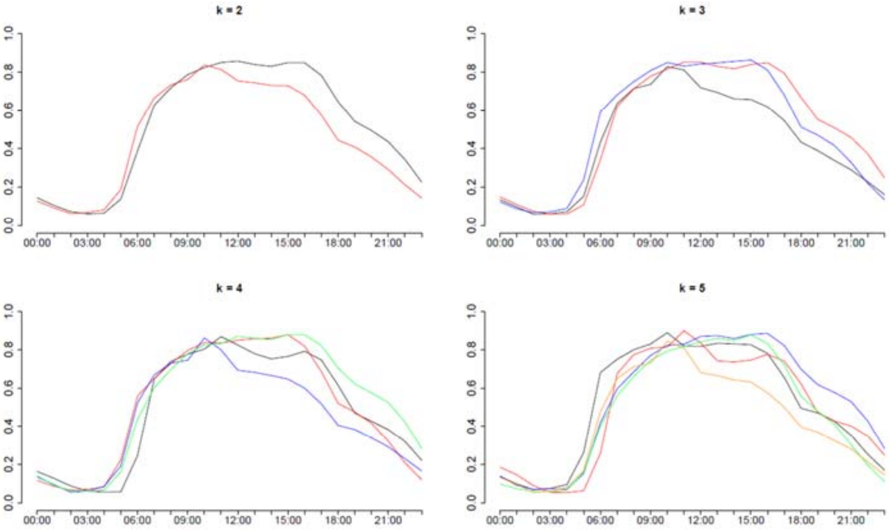

Figure 7: Clusters centroids for working days data (selected user No. 1)

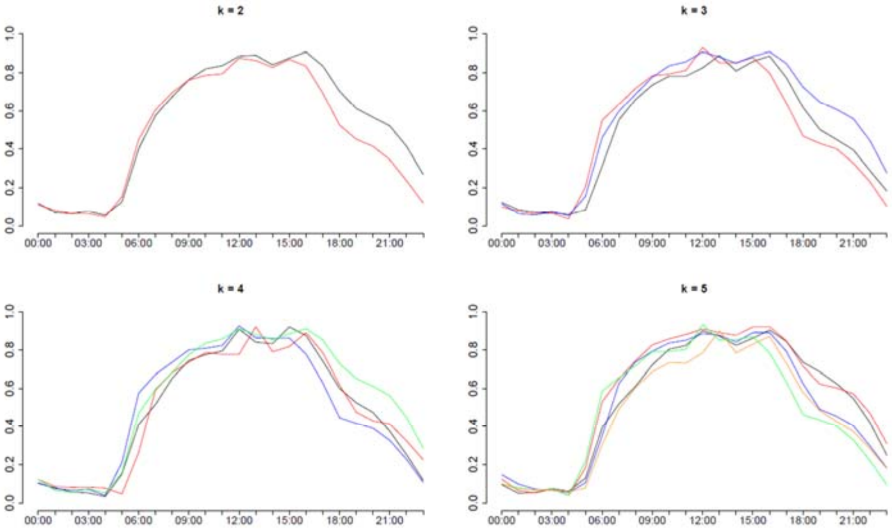

Figure 8: Clusters centroids for weekends/holidays data (selected user No. 1)

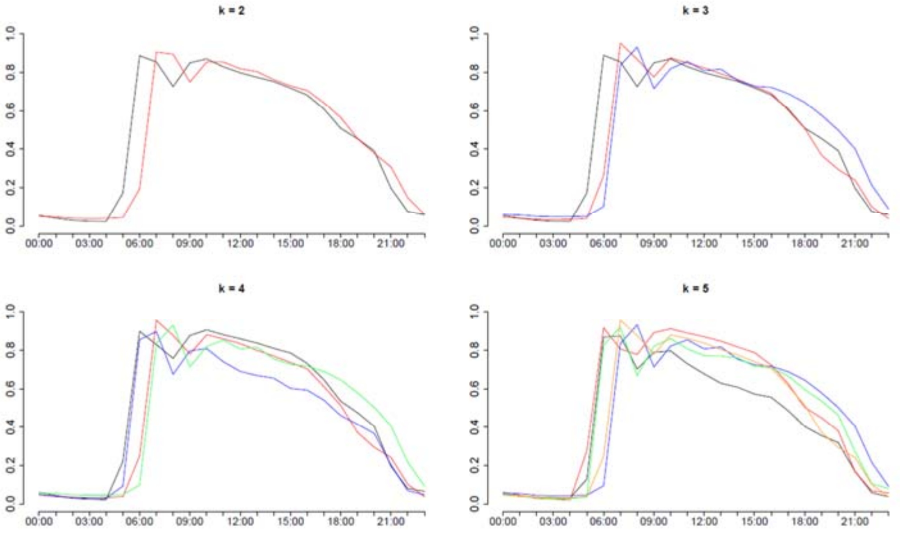

Figure 9: Clusters centroids for working days data (selected user No. 2)

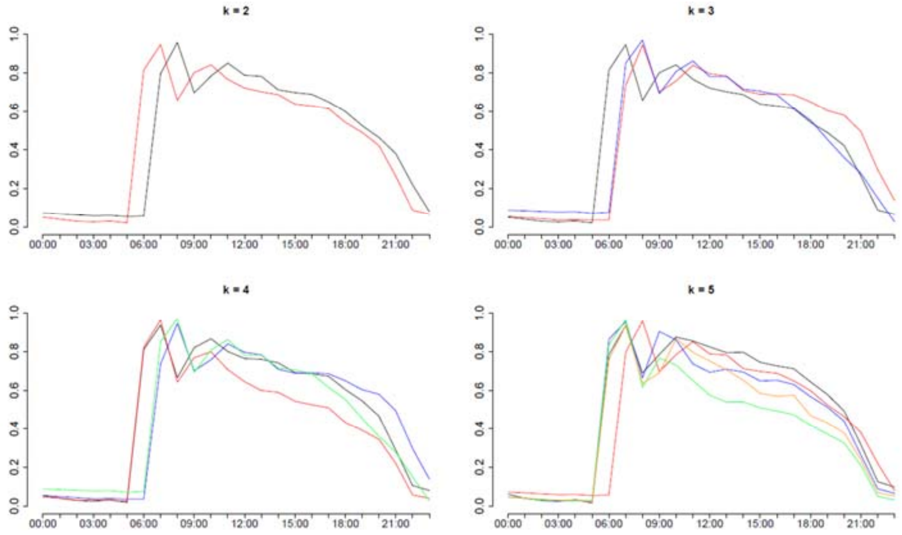

Figure 10: Clusters centroids for weekends/holidays data (selected user No. 2)

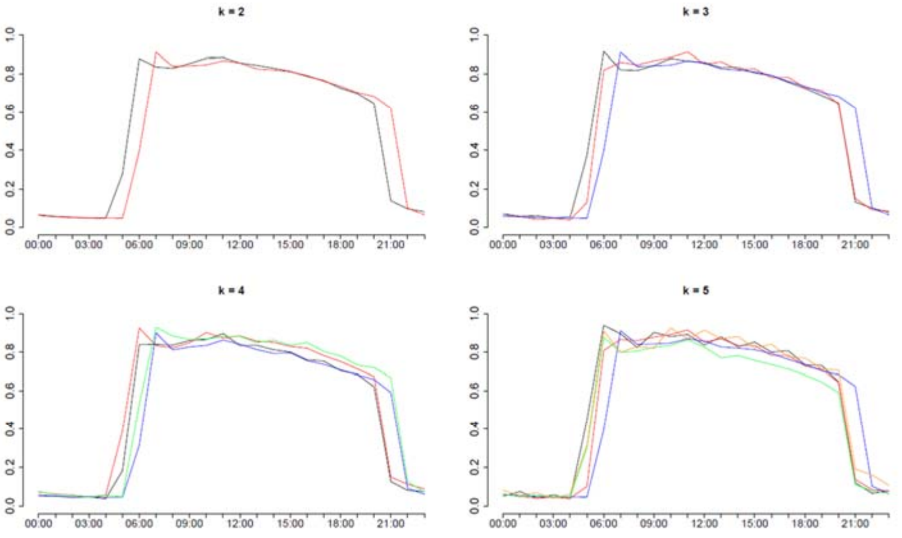

Figure 11: Clusters centroids for working days data (selected user No. 3)

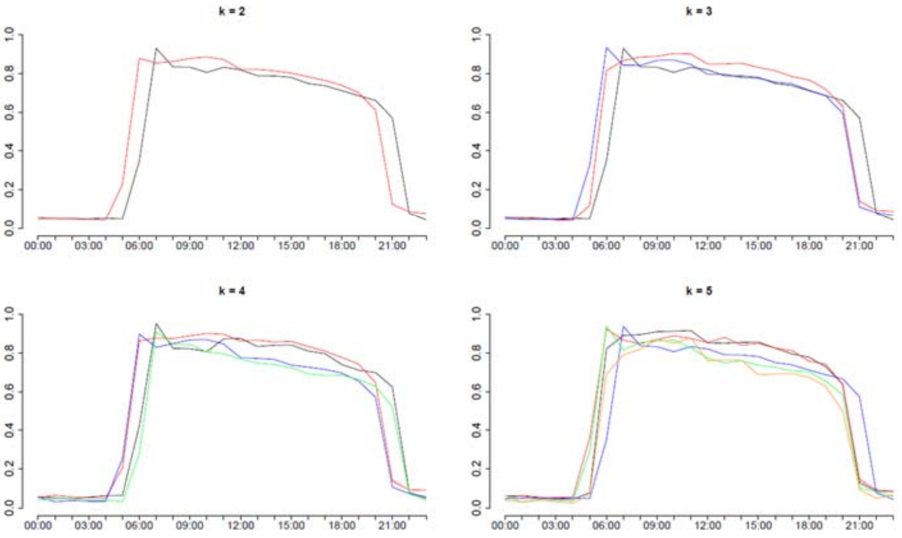

Figure 12: Clusters centroids for weekends/holidays data (selected user No. 3)

4.3 Clustering of load profiles

In order to find clusters of typical consumption patterns we use an approach described in subsection 3.4.

Results of clusters adequacy indexes are provided in Fig. 13 and Fig. 14. Analyzing results for working days data we can see that maximum value of Dunn index is at $k = 12$ while minimum value of Davies–Bouldin index occurs during cases $k = 9, k = 13$ (see Fig. 13 top left and top right). Value of cluster dispersion indicator gradually decreases while the number of clusters increases. However, from $k = 12$ change in CDI values becomes very small (Fig. 13 bottom right). Similar tendency can be seen analyzing curve of mean adequacy index (Fig. 13 bottom left). In summary, clusters adequacy indexes indicate that values $k = 9{,}12{,}13$ are the most appropriate. In order to select the number of clusters for further investigation, we analyze both the quantitative characteristics and the graphical display of results.

The comparison of cluster centroids (between cases $k = 9{,}12{,}13$, see Fig. 17, 18,15) shows that by increasing the number of clusters qualitatively different electricity consumption tendencies can be seen. Black lines depict initial tendencies, yellow lines – tendencies that have changed and green lines – new tendencies, that have not been seen before. For example, in case when $k = 13$ tendencies observed in clusters no. 5 and no. 7 have not been identified in other cases ($k = 9{,}12$). Based on these results we identified 13 different electricity consumption patterns for workings days data.

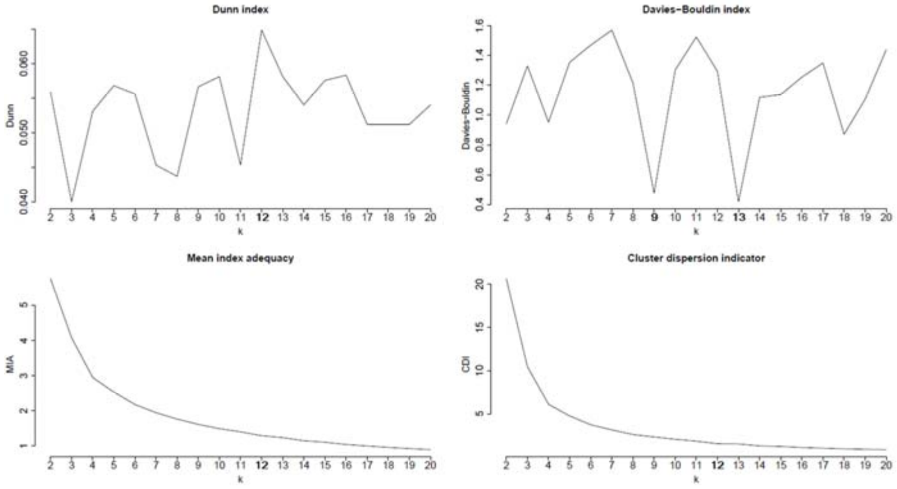

Figure 13: Clusters adequacy indices for working days data

We perform the same analysis with weekends/holidays data. In this case the results show that global maximum for Dunn index is reached, when $k = 17$ (see Fig. 14 top left). Davies–Bouldin index (Fig. 14 top right) reaches global minimum value when $k = 13$, while Dunn index at this point has value of local maximum. Value $k = 8$ indicates local minimum value for Davies–Bouldin index. Analyzing MIA and CDI curves (Fig. 14 bottom left and bottom right) we can see that from the point $k = 13$ changes in the criteria values become very small. These results show that the most appropriate values of k are $k = 8{,}13{,}17$. The particular number of clusters we select based on graphical analysis of consumption patterns (see Fig. 16, 19, 20). In these figures black lines show initial tendencies, green – new tendencies and yellow – tendencies that have change because of different number of clusters. The comparison of cluster centroids (between cases $k = 8,13,17$) shows that $k = 13$ is the optimum number of consumption patterns: in the case of $k = 8$ part of tendencies (clusters no. 10 and no. 12, Fig. 16) is lost while for $k = 17$ several tendencies are almost indistinguishable (clusters no. 3 and no.7, clusters no. 7 and no. 13, Fig. 20).

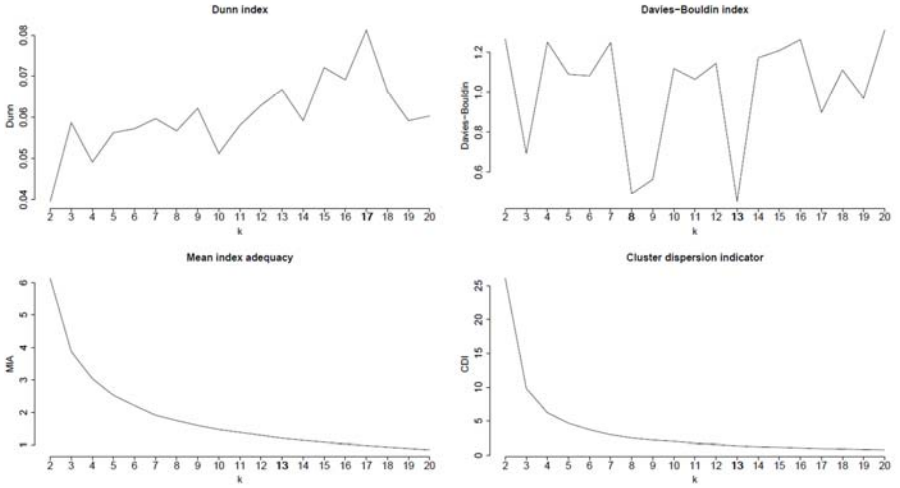

Figure 14: Clusters adequacy indices for weekends/holidays data

Clustering experiments show that 13 different customer types (profiles) can be identified for weekends/holidays and 13 for working days. Graphical partitioning results are provided in Fig. 15 and Fig. 16. Analyzing results, we can see that most of the working days patterns also occur during weekends and holidays (e.g. 3th cluster of working days and 8th cluster of weekends/holidays, 12th cluster of working days and 1th cluster of weekends/holidays), but weekends and holidays have some distinctive tendencies (e.g. 2th and 11th clusters) that are not common for working days.

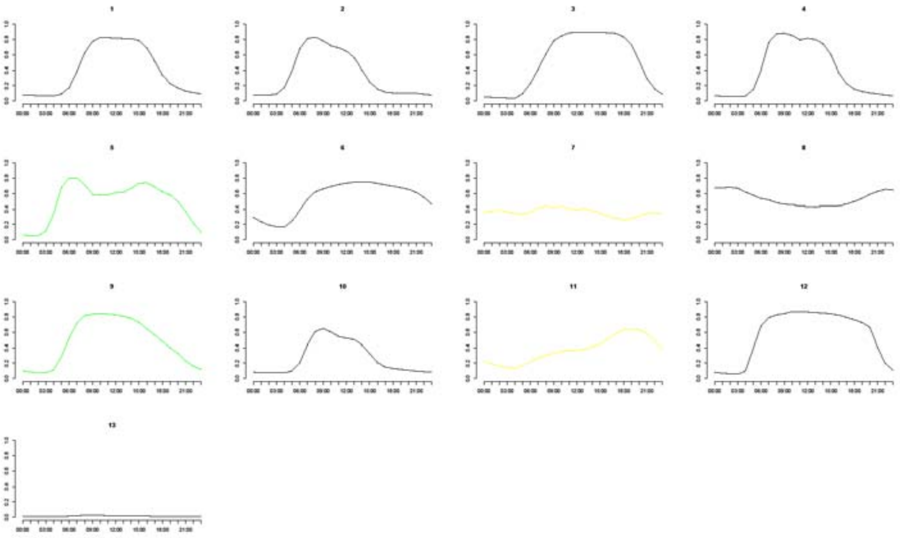

Figure 15: Electricity consumption patterns during working days

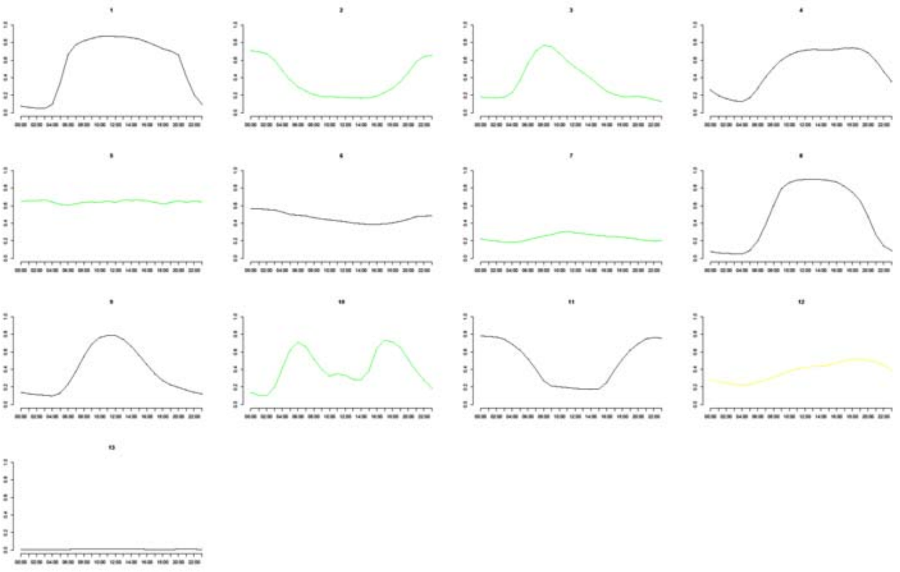

Figure 16: Electricity consumption patterns during weekends/holidays

4.4 Merging of the clustering results

Based on the approach described in subsection 3.5 we combine working days and weekends/holidays clustering results. We identify all possible combinations of working days and weekends patterns and evaluate the frequency of each combination.

Analyzing results (see table 1) we can see that some combinations are very rare[4]. These cases can be analyzed as special ones or be assigned to the one of the larger combinations based on the specific profiling task.

n w	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0	7	1	**21**	1	0	0	13	0	0	**87**	0
2	6	13	1	**25**	4	5	**16**	14	6	**21**	6	3	1
3	3	**27**	0	**20**	0	1	3	0	7	5	0	2	0
4	8	1	8	15	3	**53**	2	0	**17**	3	5	7	0
5	1	6	0	0	1	8	2	**35**	1	3	1	2	0
6	14	**29**	1	**35**	3	12	**27**	**29**	14	**25**	5	9	5
7	9	11	1	13	1	3	12	0	6	**38**	3	0	5
8	2	0	**60**	3	1	0	0	0	5	0	0	12	0
9	**41**	4	3	**29**	0	2	2	0	14	3	4	1	0
10	0	5	0	0	10	0	2	1	1	1	2	0	0
11	11	**27**	0	**31**	0	2	7	**26**	2	**25**	4	3	0
12	14	8	2	**28**	1	10	12	4	5	**18**	**23**	0	1
13	0	1	0	2	0	1	1	1	0	8	0	0	**151**

Table 1: Combinations of working days (w) and holidays (n) clusters

Identification of a threshold value for the minimum size of profile (i.e. a minimum number of users in profile) is a marketing problem. In this example we will consider that all combinations $f_{ij}, i,j = 1,\dots,13$ with frequency higher than 15 can be identified as distinct electricity usage profiles. Therefore, the proposed profiles are marked in bold font in table 1. The rest of the combinations can be analyzed as special cases or be assigned to one of the previously defined profiles using one of the methods proposed in subsection 3.5. Since the selection of an appropriate approach is a marketing problem it will not be further analyzed.

5 Conclusions

We have presented an approach for the grouping of electricity customers based on the similarities of their (hourly) consumption patterns. We proposed a several stages methodology, i.e. using K-means to obtain load profiles for users and to partition these profiles into homogeneous clusters.

Experiments show that proposed approach is able to provide well-separated clusters clearly representing the behavior of electricity users. Moreover, method leaves to the supplier possibility of defining the number of profiles by adjusting the threshold value for the minimum size of profile. Furthermore, algorithm enables the possibility to choose how to design profiles for load patterns with uncommon behavior. Even though proposed methodology was tested using only one dataset we expect the approach to generalize to other datasets properly. Proposed methodology is

[4] In this case the combination of load profiles is defined as a rare if it is ten times smaller than the largest combination. However, this value can be adjusted based on marketing decisions and the aim of consumption analysis.

based on K-means algorithm. This technique stands out for being considered one of the ten most influent algorithms in data mining. Such influence is due to its simplicity, scalability, and easy adaptation to different domains [Kumar, 09].

Experiments revealed that in this case all users have only one consumption pattern during weekends/holidays and one during working days. However, this assumption may not always be valid. For this reason, in the future we are going to examine more complicated cases:

1. users having more than one consumption pattern,
2. users having more typical behaviors than just working and non - working days.

For this reason, we are considering the possibility to use soft assignment k-means algorithm which is an extension of k-means where each data point can be a member of multiple clusters with a membership value. Moreover, clustering adequacy indexes (e. g. *MIA*, *CDI*) require a biased exert knowledge. In the future we are planning to develop more objective methodology for selection of parameter k. We are considering usage of MapReduce algorithm [Naldi, 15] in order to evolve clusters without specifying k-means parameters. Furthermore, future plans include methods for dealing with missing or incomplete data. We are interested in several possible solutions:

1. Partial distance strategy (PDS) – distance between two objects is calculated using only observed values.
2. Optimal completion strategy (OCS) – missing values are regarded as additional parameters to be optimized during clustering.
3. Nearest prototype strategy (NPS) – modification of OCS, in which missing values are imputed considering only the nearest prototype.

Acknowledgments

We thank Vytautas Mickevičius[5] for help developing prototype system, and Aleksandras Kazimianec[6] and Gediminas Trakas[7] for domain expertise. This work is partially supported by the XJTLU Key Programme Special Fund (KSF-P-02), AI University Research Centre (AI-URC), Xi’an Jiaotong-Liverpool University, China

References

[Aickelin, 11] U. Aickelin, I. Dent and T. Rodden. Application of a clustering framework to UK domestic electricity data. In UKCI 2011, the 11th Annual Workshop on Computational Intelligence, Manchester, pages 161–166, 2011.

[Aickelin, 12] U. Aickelin, I. Dent, T. Craigy and T. Roddenz. An approach for assessing clustering of households by electricity usage. In proceeding of: UKCI 2012, 12th Workshop on Computational Intelligence, 2012.

[Babos, 02] A. Babos, F. Kovacs, C. Legany. Cluster validity measurement techniques. Technical report, Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Budapest, Hungary, 2002.

[5] Vytautas Magnus University

[6] Binar Solutions, UAB

[7] Gera Solutions, UAB

[Bibby, 79] J. M. Bibby, K. V. Mardia, J. T. Kent. Multivariate Analysis. Academic Press, 1979.

[Examples] Examples of time series and their periodograms https://www.dropbox.com/s/5ob0sjmm3d6otj8/Examples.pdf?dl=0.

[Flath, 12] Ch. Flath, D. Nicolay, T. Conte, C. van Dinther, and L. Filipova-Neumann. Cluster analysis of smart metering data - an implementation in practice. Business & Information Systems Engineering, 4(1):31–39, 2012.

[Han, 12] J. Han and M. Kamber. Data mining concepts and techniques. Morgan Kaufmann Publishers, 2 edition, 2012.

[Kumar, 09] V. Kumar, X. Wu. The Top Ten Algorithms in Data Mining. Chapman & Hall/CRC, 1st edition, 2009.

[Liu, 15] L. Liu. Cluster analysis of electrical behavior. Journal of Computer and communications, 3:88–93, 2015.

[Naldi, 15] M. C. Naldi, G. V. de Oliveira. Scalable fast evolutionary k-means clustering. In proceeding of: 2015 Brazilian Conference on Intelligent Systems, pages 74–79. IEEE.

[Poggi, 15] J. M. Poggi, E. Devijver, Y. Goude. Clustering electricity consumers using high-dimensional regression mixture models, 2015.

[Rasanen, 10] T. Rasanen, D. Voukantsis, H. Niska, K. Karatzas, and M. Kolehmainen. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Applied Energy, 87:3538-3545, 2010.

[Scargle, 82] J.D. Scargle. Studies in astronomical time series analysis II: Statistical aspects of spectral analysis of unevenly spaced data. Astrophysical Journal, 263:835–853, 1982.

[Wong, 12] J. Wong, B. A. Smith and R. Rajagopal. A simple way to use interval data to segment residential customers for energy effciency and demand response program targeting. ACEEE Summer Study on Energy Efficiency in Buildings, pages 374–386, 2012.

Appendices

A. Cluster centroids for working days set

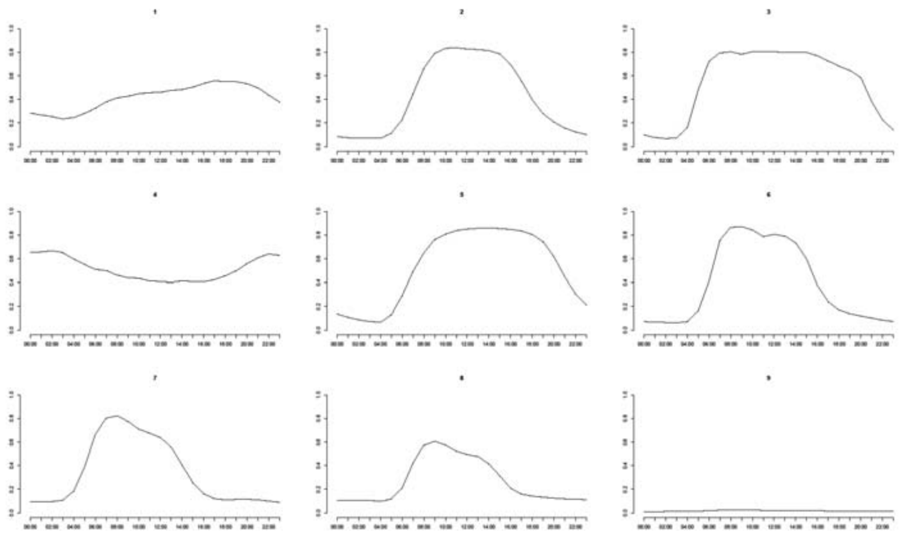

Figure 17: Electricity consumption patterns during working days when k=9

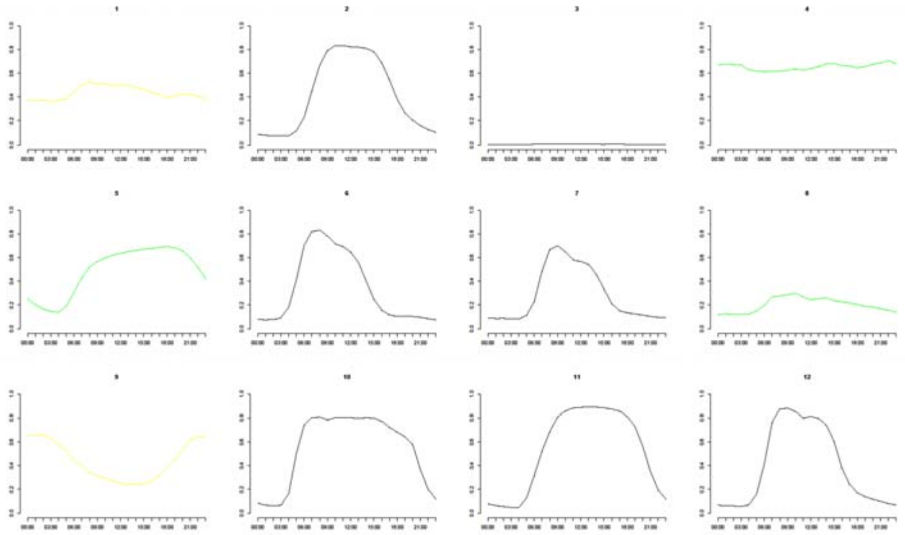

Figure 18: Electricity consumption patterns during working days when k=12

B. Cluster centroids for non - working days set

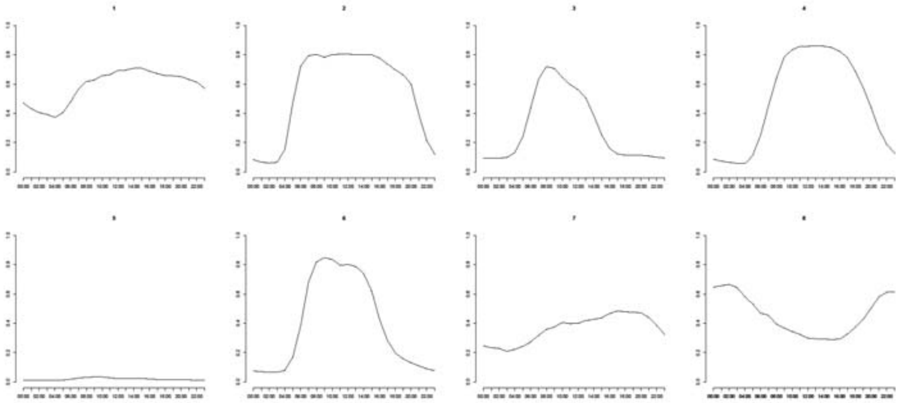

Figure 19: Electricity consumption patterns during non - working days when k=8

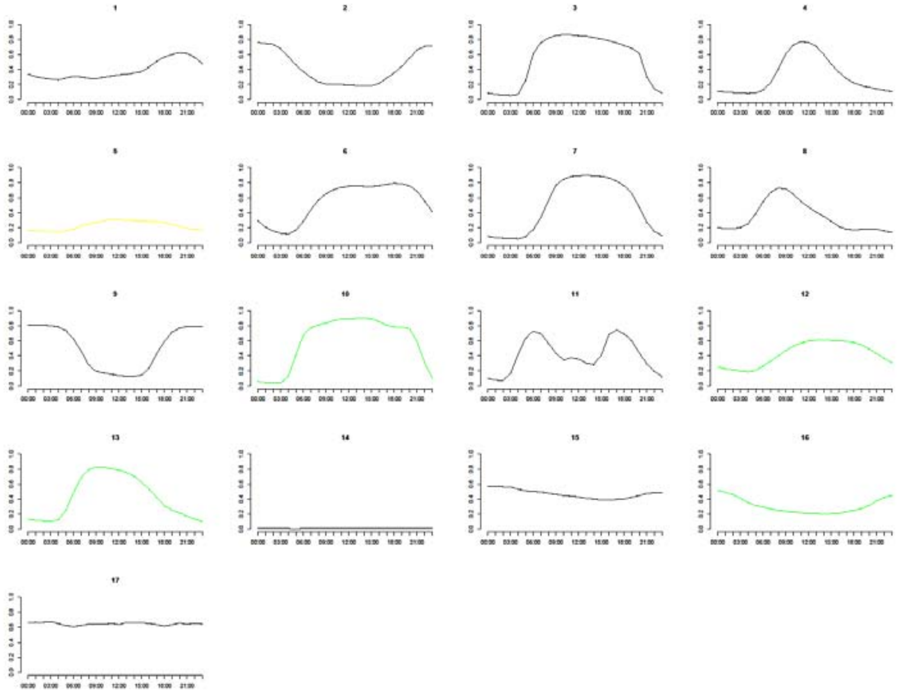

Figure 20: Electricity consumption patterns during non - working when k=17