Organizing, Visualizing and Understanding Households Electricity Consumption Data through Clustering Analysis

Christian Nordahl, Veselka Boeva, Håkan Grahn, and Marie Persson Netz
Blekinge Institute of Technology, Karlskrona, Sweden
{christian.nordahl, veselka.boeva, hakan.grahn, marie.netz}@bth.se

Abstract

We propose a cluster analysis approach for organizing, visualizing and understanding households' electricity consumption data. We initially partition the consumption data into a number of clusters with similar daily electricity consumption profiles. The centroids of each cluster can be seen as representative signatures of a household's electricity consumption behaviors. We evaluate the proposed approach by conducting a number of experiments on electricity consumption data of ten selected households. Our results show that the approach is suitable for data analysis, understanding and creating electricity consumption behavior models.

1 Introduction and Related Work

With the adoption of smart meters in the electrical power grids, we have the opportunity to collect high resolution electricity consumption data remotely on a household level. Today, collection usually takes place every 30-60 minutes. This type of data can be used to get insight into the residents' habits and activities, with low impact and intrusion of the residents' privacy.

Most current research related to household electricity consumption have mainly revolved around creating consumer profiles by clustering households together [Chen *et al.*, 2015], or comparing different households to determine and predict abnormal consumption patterns [Chalmers *et al.*, 2015]. There has been less attention and focus on analyzing, understanding and modeling of the electricity consumption behavior of individual households.

We may detect abnormalities and changes of residents' behavior through analyzing their daily household electricity consumption. For example, dementia, and other neurodegenerative diseases, cause changes in the behavior of the individual in different ways, e.g., they can provoke insomnia, apathy, restlessness etc. [Mega *et al.*, 1996]. We believe that changes like these, in the individual's daily behavior, can be caught by his/her electricity consumption activities.

In this paper, we present and evaluate a cluster analysis approach for organizing, visualizing and understanding electricity consumption data. Our aim is to study the possibility of using the knowledge discovered by such analysis for creating consumption behavior signatures on a household level. The long-term goal is to investigate whether the created signatures can be used for identifying abnormal behavior in daily life and apply this outlier detection model in health care applications, e.g., for monitoring early stages of dementia or other neurodegenerative diseases. The developed consumption signatures can be considered as predefined activities and can be used for detecting abnormal consumption patterns in order to notify the environment (relatives and health care professionals) if early signs of dementia occurs repeatedly at home.

2 Methods

2.1 Data Segmentation and Standardization

Our electricity consumption data is collected with one minute resolution, but we aggregate it into one hour segments. Thus, a daily profile consists of 24 data points, each representing one hour of electricity consumption (in *kWh*), that we can analyze as a time series.

Before performing clustering and further analysis, we standardize the time series profiles using z-standardization. Z-standardization is suitable when the general shape of the time serie is more important than the amplitudes at the different time points. Therefore, the electricity consumption profile of each day is adjusted by subtracting the profile's mean and dividing with the profile's standard deviation.

2.2 Clustering Algorithms

Three partitioning algorithms are commonly used for data analysis to divide the data objects into k disjoint clusters [MacQueen and others, 1967]: k-means, k-medians and k-medoids clustering. The three partitioning methods differ in how the cluster center is defined. In k-means clustering, the cluster center is defined as the mean data vector averaged over all objects in the cluster. In k-medians, the median is calculated for each dimension in the data vector to create the centroid. Finally, in k-medoids clustering, which is a robust version of the k-means, the cluster center is defined as the object with the smallest sum of distances to all other objects in the cluster, i.e., the most centrally located point in a given cluster.

We have used k-medoids in our experiments, because we believe that having an actual consumption profile as the cluster's centroid (medoid) is more representative of the consumption behavior compared to creating a synthetic centroid. k-medoids has been implemented in accordance to its description given in [Bauckhage, 2015].

2.3 Distance Metrics

In our experiments, we use and compare two different distance metrics; Euclidean distance (ED) and Dynamic Time Warping (DTW). ED calculates the distance between two vectors points by point. DTW is a distance metric, that is specifically designed for time series data [Sakoe and Chiba, 1978]. DTW aims at aligning two time vectors by warping the time axis iteratively until an optimal match (according to a suitable metric) between the two vectors is found. Thus, DTW is a much more robust distance measure for time series than classical distance metrics, such as ED, since it allows similar shapes to match even if they are out of phase in the time axis.

In our experiments we have used FastDTW [Salvador and Chan, 2007], which is a computationally optimized implementation of the standard DTW algorithm.

2.4 Cluster Validation Measures

Cluster validation techniques are designed to find the partitioning that best fits the underlying data [Halkidi *et al.*, 2001]. The general idea is to choose the best clustering scheme, of a set of defined schemes, according to a pre-specified criterion.

The partitioning algorithms contain the number of clusters (k) as a parameter and their major drawback is the lack of prior knowledge for that number to construct. Unfortunately, determining a correct, or suitable, k is a difficult problem in a real data set. For such cases, researchers usually generate clustering results for different numbers of clusters, and subsequently assess the quality of the obtained clustering solutions. For example, internal cluster validation measures such as Silhouette Index (SI) [Rousseeuw, 1987] and Connectivity [Handl *et al.*, 2005] can be used as validity indices to identify the best clustering scheme.

3 Results and Discussion

3.1 Data

We gathered electricity consumption data from 9909 households, collected with a 1 minute interval, all of which are anonymous. The households' data is preprocessed as explained in Section 2.1. However, before performing z-standardization of each daily consumption profile, we remove all profiles which have more than 10% missing values in total or have more than 20 consecutive missing values. The missing values of the remaining daily consumption profiles are imputed by using linear interpolation. Notice that whenever there are missing values present, for any household, they usually last for entire days causing most of them to be excluded.

We evaluate our data analyzing approach by conducting experiments on 10 households. These 10 households are chosen because they contain the highest number of daily profiles after the pre-processing stage. In this paper, we present and discuss results obtained for one of these households, which is representative for all chosen households.

3.2 Estimation of the Number of Clusters

We run the clustering algorithm (k-medoids) with two different distance metrics (ED and DTW) for all values of k between 2 and 9. Then, we plot the values of the Silhouette index and the Connectivity score obtained by each k, see Figure 1. We use the SI and Connectivity measures as validity indices to identify the best partitioning scheme. We search for the values of k at which a significant local change in value of the index occurs [Halkidi *et al.*, 2001]. These values are different for the considered validity indices.

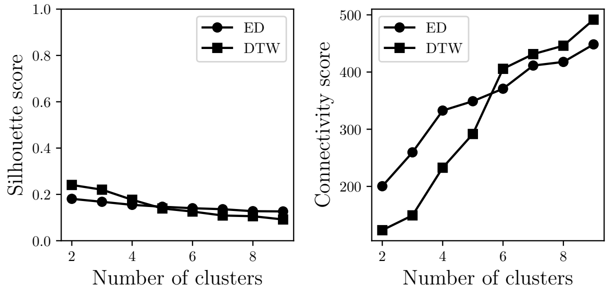

Figure 1: Silhouette index (left) and Connectivity (right) scores generated by k-medoids on the household consumption data set using ED and DTW distance metrics for different number of clusters.

The SI provides a near horizontal line, and the Connectivity measure continuously increases with the number of clusters. This can be explained by the fact that many of the profiles overlap in their idle state, e.g., when the residents are not at home, which causes both measures to find neighbors in other clusters in close proximity. Neither metric give a clear indication of what is the appropriate k. It is, however, interesting to observe that the scores generated by both validity measures using DTW distance are higher than ones obtained by ED for $k \leq 5$ and vice versa when $k > 5$.

Due to the inconclusiveness of the used validation metrics, we visually inspect the produced clusters. We determine that $k = 7$ for ED and $k = 6$ for DTW produced reasonable clusters for this household.

3.3 Clustering Analysis

Figure 2 and Figure 3 show the clusters produced by k-medoids using two different distance metrics: ED and DTW, respectively. In both figures the gray lines are the z-standardized daily consumption profiles assigned to the clusters and the black lines are the clusters' medoids. The figures depict the results generated by k-medoids for $k = 7$ with ED, and $k = 6$ with DTW, respectively.

Consumption Behavior Signatures

The clusters' medoids presented on Figure 2 and Figure 3 can be considered as signatures of electricity consumption habits of the household residents. For example, as one can notice in Figure 2 the signatures of clusters 3, 4, 5 and 6 present clearly defined consumption activities, supported by a high number of daily profiles assigned to these clusters. The latter is also valid for clusters 0, 2, and 3 in Figure 3.

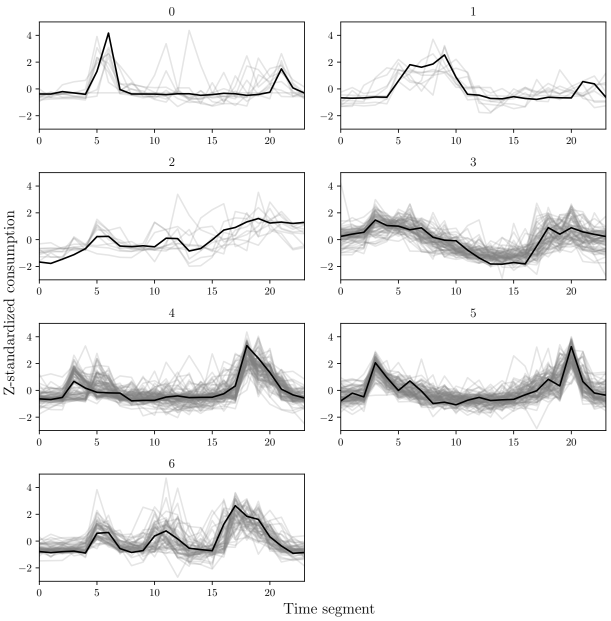

Figure 2: Clusters generated by k-medoids for $k = 7$ with ED.

Table 1 and Table 2 show how the weekdays are distributed within each cluster and the ratios between working days and weekends for ED and DTW, respectively. The ratio is calculated by the mean of working days divided by the mean of weekend days. ∞ occurs when the cluster contains no weekends, due to a division by 0.

We can observe in Table 1 that cluster 6 contains a clear majority of weekends, while clusters 3, 4, and 5 have mostly working days. This is supported by their respective signatures given in Figure 2. Namely, we can see a typical electricity consumption behavior of employed residents for all these clusters. For example, there are clear consumption peaks in the morning and evening, while for the remainder of the day the electricity consumption remains at an idle state. Cluster 6, on the other hand, has an additional peak of consumption in the middle of the day, showing that the household residents are active at home. It is also interesting to notice that the signatures of clusters 4 and 5 (see Figure 2) are very similar, which can be an indication that these could be merged into a single cluster. This is also supported by the results in Table 1, because these are the largest clusters and also both clusters present a typical working day consumption behavior (see the calculated ratio scores).

The similar trends in terms of the distribution of working days and weekends can be seen in Table 2. The largest clusters (0 and 2) contain a clear majority of working days, while cluster 3 (the third largest) has mostly weekend days. The signatures produced by these clusters (see Figure 3) are similar to the signatures of clusters 4 & 5, and 3 & 6, respectively, as shown in Figure 2. Namely, clusters 0 and 2 show the typical working day consumption behavior, and cluster 3 has the same additional peak in the middle of the day.

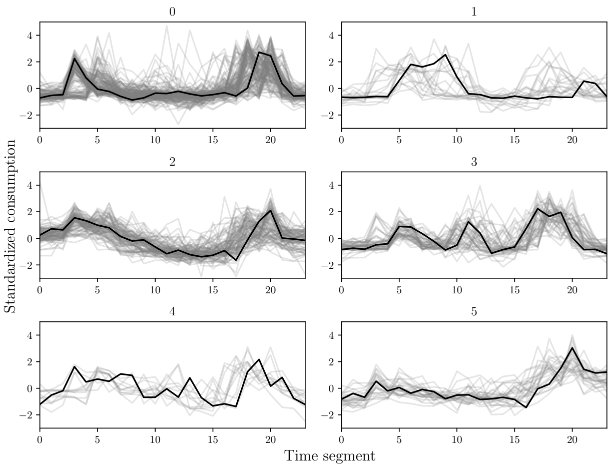

Figure 3: Clusters generated by k-medoids $k = 6$ with DTW.

Cluster	0	1	2	3	4	5	6
# days	11	9	11	72	94	92	58
Ratio	0.229	0.2	0.48	1.4	1.95	∞	0.083

Table 1: Number of days and the ratio between working days and weekends in each cluster generated by k-medoids using ED.

Cluster	0	1	2	3	4	5
# days	142	20	95	50	12	28
Ratio	2.305	0.6	1.711	0.141	0.286	1.2

Table 2: Number of days and the ratio between working days and weekends in each cluster generated by k-medoids using DTW.

The remaining clusters are much smaller in comparison to the others for both distance metrics (ED and DTW). The smaller clusters produced by k-medoids with ED (clusters 0, 1 and 2) contain mostly weekends. This is also valid for two smaller clusters (clusters 1 and 4) generated by using DTW. This observation is supported by the fact that people usually have more diverse activities during weekend days.

Context-based Signatures

In addition to the consumption behavior signatures, we generate context-based signatures. We can model the context by presenting each cluster by a vector of relative contributions of the different weekdays to it. These context-based signatures for the clusters generated by k-medoids and discussed in the previous section are shown in Figure 4 and Figure 5.

By adding this type of context to our consumption behavior model, we can conduct further analysis and extract additional knowledge about the electricity consumption habits of the household residents. We can also use these context-based signatures for further refinement of the developed consumption behavior model. For example, clusters 0 and 2 in Figure 5 have similar context-based signatures. In addition, their consumption behavior signatures are also very similar. This may be interpreted as a suggestion for merging these clusters into a single one. It is also interesting to compare the context-based signatures generated by two distance metrics. We can observe that the signatures of the clusters generated by using

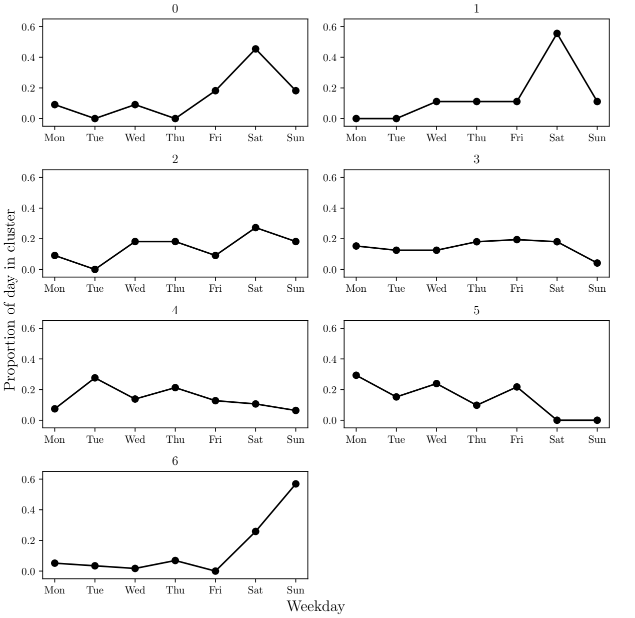

Figure 4: Context-based signatures of the clusters generated by k-medoids for $k = 7$ with ED.

DTW are more monotonic. Namely, the working days are more uniformly distributed between these clusters.

4 Conclusions and Future Work

In this study, we have proposed a cluster-based approach for organizing, visualizing and understanding household's electricity consumption data. We have evaluated the proposed approach on 10 households and the initial results show that the produced clusters well reflect different household electricity consumption behaviors. We have shown how the clusters' context can be modelled and we also discussed how the created context-based signatures can be used for further understanding and refinement of the developed consumption model. Our long-term aim is to use this approach for building an electricity consumption model that can be used for detecting abnormal behavior of elderly individuals, e.g., ones with early stages of dementia or other neurodegenerative diseases.

As future work, we initially plan to apply and evaluate the proposed approach on households' electricity consumption data that are representative for elderly individuals. A process for collecting electricity consumption data from a number of households has been recently initiated. Our intention during this ongoing process is to be in continuous contacts with the residents to be able to correctly label the daily consumption profiles. Further, we will also evaluate our data analyzing approach using other distance metrics and clustering algorithms, e.g., distance metrics that have been developed for shape based clustering, including ShapeDTW and Shape-based Distance in conjunction with k-shape.

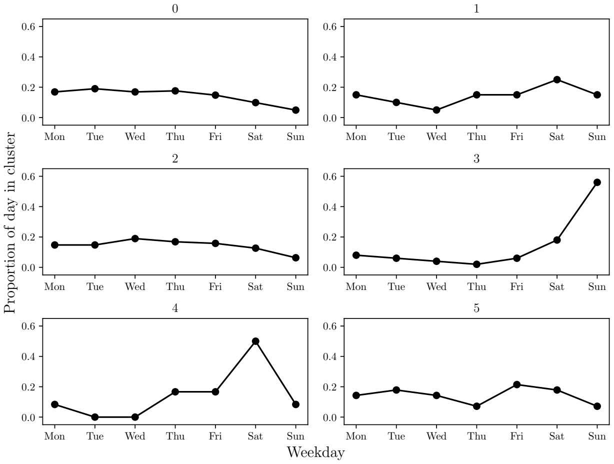

Figure 5: Context-based signatures of the clusters generated by k-medoids for $k = 6$ with DTW

References

[Bauckhage, 2015] C. Bauckhage. Numpy/scipy recipes for data science: k-medoids clustering. *researchgate.net, Feb*, 2015.

[Chalmers *et al.*, 2015] C. Chalmers, W. Hurst, M. Mackay, and P. Fergus. Profiling users in the smart grid. In *Seventh Int'l Conf. on Emerging Networks and Systems Intelligence*, 2015.

[Chen *et al.*, 2015] T. Chen, A. Mutanen, P. Järventausta, and H. Koivisto. Change detection of electric customer behavior based on AMR measurements. In *PowerTech, 2015 IEEE Eindhoven*, pages 1–6, 2015.

[Halkidi *et al.*, 2001] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *J. of Intelligent Information Systems*, 17(2-3):107–145, 2001.

[Handl *et al.*, 2005] J. Handl, J. Knowles, and D.B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

[MacQueen and others, 1967] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability*, pages 281–297, 1967.

[Mega *et al.*, 1996] M.S. Mega, J.L. Cummings, T. Fiorello, and J. Gornbein. The spectrum of behavioral changes in alzheimer's disease. *Neurology*, 46(1):130–135, 1996.

[Rousseeuw, 1987] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. and Applied Mathematics*, 20:53–65, 1987.

[Sakoe and Chiba, 1978] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[Salvador and Chan, 2007] S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.