Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns

Hông-Ân Cao
Department of Computer Science
ETH Zürich
8092 Zürich, Switzerland
Email: hong-an.cao@inf.ethz.ch

Christian Beckel
Department of Computer Science
ETH Zürich
8092 Zürich, Switzerland
Email: beckel@inf.ethz.ch

Thorsten Staake
Energy Efficient Systems
University of Bamberg
96047 Bamberg, Germany
Email: thorsten.staake@uni-bamberg.de

***Abstract*—Elaborating demand side management strategies is crucial for integrating electricity from renewable sources into the electrical grid. Though future demand side will largely depend on an automatic control of larger loads, it is also widely agreed upon that consumer behavior will play an important role as well - be it by purchasing respective automation techniques or by shifting the use of appliances to other times of the day. Doing so, it becomes possible to select households that offer sufficient load shifting potential, and to overcome undirected and thus, expensive campaigns. To our knowledge, this perspective is still under-researched, especially when it comes to clustering methods on load consumption data with a focus on peak detection accuracy to provide customer segmentation.**

Using the data collected in the Irish CER dataset, which contains readings for more than 4000 residential customers over a period of 18 months at 30-minute intervals, we show that the whole clustering of the time series, with a few adaptations on the usage of the K-Means algorithm, provides better clustering results without sacrificing practical feasibility. Characteristic load profiles allow us to segment the customers, address groups of households with similar consumption patterns and determine on the fly the cluster membership of a given load curve. This will support decision making regarding the investments in load shifting campaigns to prevent over or under-dimensioning linked to peak energy demand.

I. Introduction

As the generation of electricity from renewable resources does not fully rely upon a previously defined and arbitrary schedule, but is the result of varying environmental parameters, the required flexibility to balance supply and demand will increasingly be achieved by managing the demand side of the grid. This will require a more thorough appreciation of the network flow and usage, and contrasts with the current set-up, where synthetic load profiles are commonly used to provision for energy, although they constitute an average profile for all households. More fine-grained information about the specific consumption patterns could allow for a better understanding of when and which customers are responsible for the peak-time energy consumption, which is costly for the energy providers.

While extensive work has been carried out on producing an estimation of the load consumption, we are focusing on identifying the characteristic load profiles. The novelty resides in the fact that although clustering methods have been tested on smart-metering data, they were mostly intended as an exploratory phase or as a proof of concept that the data can be segmented. For this reason, evaluation means of the obtained clusters still need to be defined in order to be applicable for clearly defined use cases.

We see three major advantages that relate to (i) providing detailed insights about household load curves in general (ii) being able to identify “hurtful” households, which helps to focus the cost of mitigation to the relevant ones (iii) having a means that can assist in determining the customer value up to the point where tariffs may depend on the load curve, even in the household segment. A non exhaustive list would comprehend measures such as sending prompts, extended information on utility bills, behavioral cues e.g., to collect bonus points for a desired change of load profile type, enabling energy consulting teams to pre-select households that are given priority for automatic load shifting measures or evaluating the effects of load shifting campaigns in a very focused way.

The work is of special interest as it can be implemented without hardware investments beyond off the shelf deployment of smart-metering infrastructures, using well known but specifically adapted clustering techniques. It relies on a new approach to select the appropriate parameters and establishing characteristic cluster profiles as references to determine cluster membership on the fly.

The paper is structured as follows. We review the related work in Section II. Then, we present the methodology to build the clustering framework in Section III and discuss the results in Section IV. We provide insights on possible applications and research tracks in Section V.

II. Related Work

Load forecasting has been explored with the aim of predicting the load to be provisioned based on historic data by [1]. To our knowledge, [2] also investigated the Irish CER dataset, but the work focuses on the segmentation of households and relies more on survey data, which relates to a classification task. While clustering has been considered by [3]–[9], an evaluation of the "quality" of the obtained clusters has not yet been undertaken: this relates to the choice of the clustering algorithm, the distance measure that is examined and an analysis and discussion of shapes of the obtained characteristic load profiles. [10] has provided a very thorough analysis and comparison of different clustering techniques. The work of [11] shows that care must be taken when mining data from time series to be able to justify the claims related to the results of an empirical evaluation.

III. Methodology

We concentrate on cluster consumption patterns based on peak positions, which can be identified as hurtful moments of the day for energy providers. This would allow not only to characterize populations of customers, but also to react to the more demanding profiles. The latter can be enabled by adopting a strategy of contacting them and offering counseling or different tariffs, in order to influence their consumption behavior to fit the utility companies' goals. To target customers that are more likely to react positively to such stimuli, their selection can be supported by favoring stable behaviors over time (households that don't significantly change their time of peak consumption from one week to the other, which can be seen as stable[1]). A way for the utility companies to better provision for their network without relying only on synthetic load profiles can be foreseen, which might be aggregating the information too much and thus, be less adaptive to the specificity of the population of customers that are served.

The Irish CER dataset contains 30-minute readings of 4225 residential customers, which were collected over a period of 18 months throughout Ireland. Building the analysis with these data allows to show that the results are significant and not influenced by an ill-sampled, hence not representative enough set of households.

To achieve robust results, the first task consists in assessing the quality of the input and deciding the format of the object to be clustered. Then, we explore suitable clustering methods and the choice of parameters that can enable the identification of peaks in the load curves.

A. Data pre-processing

Overall, we focus on the shape of the curve instead of the exact amount of consumed energy. The goal is not to forecast the load at any point in time, but rather to target a set of clusters that diverge in the position of their peaks throughout the day. We review the required steps to build the objects that will be clustered.

1) Cleaning of the dataset: Best practice in data mining consists in verifying the quality of the input. Given the reports [12] and [13], we assumed the presence of potential hardware failure and discarded the data collected from the first month. 0 kWh readings were nevertheless identified throughout the span of the data collection. Their presence could be attributed to smart-metering faults[2]. Different reasons could be suggested such as blackouts (in the case where the smart-meter is self-powered, such reading could be happening). Figure 1, shows the case of one particular household with multiple 0 kWh readings. We can clearly see that this pattern does not apply only to specific weekdays, but to any day of the week. Hence, we decided to investigate the occurrence of the null measurements through histograms of their distribution. The adopted strategy relied on evaluating the proportion of incriminated consecutive measurements. When looking at the maximum length of consecutive null readings, the number of affected daily curves quickly dropped below 100 as the length of the sequence increased. This motivated the choice of a sequence of five 0 kWh records as a cut-off value for the removal of incriminated curves. This allowed the discarding of a very negligible number of curves overall (0.55% to 0.8% of all curves in the datasets listed in Table II).

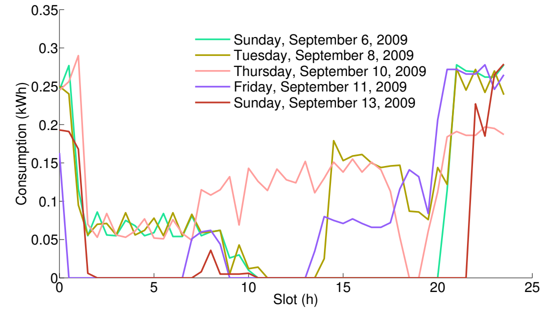

Fig. 1. Daily curves for one household with multiple consecutive 0 kWh readings. We notice that the issue is not related to one single day of the week, but over any day of the week.

The data were collected at a frequency of once every 30 minutes, providing 48 samples per day. However, this was not the case when daylight saving time (DST) was implemented. In the case of shifting to winter time, one additional hour was added to the daily readings, implying that 50 samples were recorded and when moving to summer time, one hour disappeared, thus only 46 samples were kept. This was mitigated by correcting the incriminated days and transforming the corresponding vectors of readings into regular 48-sample vectors. For this reason, the third and fourth samples on the winter DST day were discarded, since they are duplicated readings from 1 am to 2 am. Regarding the summer time, as records from 2 am to 2:59 am were missing, the average of the first hour of data was replicated.

2) Splitting the dataset: Some authors highlighted and incorporated seasonal differences in their implementation and analysis such as [14]. The dataset was divided into summer and winter data, as some seasonality effect was expected to influence the shape of profiles. For each "season", four weeks of data was used as "training data" to build the cluster profiles, while the larger corresponding sets can be designated as "test data".

[1]On the contrary, it could be argued that unstable households are of interest.

[2]0 kWh readings should not happen: http://www.ss3meteronline.co.uk/faq.html

3) Focusing on the load curves shapes: We decided to evaluate weekday patterns by averaging weekly data from Monday to Friday and removing special days such as public holidays. The variation from one weekday to the other was not significant enough to build separate patterns for each day as expressed in [14].

The interest being primarily the overall shape of the load profiles, we considered the effect of a Wiener filter to remove the oscillations, which, in the framework set-up, are less relevant than the most prominent peaks. For this purpose, we selected different smoothing windows.

Similarly, we examined different normalization and scaling methods that are presented in [15]. As described in [15], normalizing each curve by dividing it by its maximum value not only preserves the shape of the curve, but also provides a scaling between 0 and 1 of all measurements. This preserves the relative variability between each reading and renders each object independent from each other and from the dataset (this is not the case if column-wise modifications are applied on the raw data for example). This was further consolidated by comparing the outcome of the clustering using the different normalizing techniques, which provided the most differentiated cluster profiles, i.e. cluster separation.

We performed formatting on really low consumption load curves, to deal with cases where the dwelling is left inhabited and which are expected to only present a base-load consumption. To determine the threshold to separate base-load/standby consumption from "real" user triggered consumption patterns, we plotted a histogram of the distribution of the average weekly consumption. For bins of 0.025 kWh from 0.025 kWh to 3 kWh (the maximum), we determined that for all average weekly figures below 0.125 kWh[3] the order of magnitude of the affected curves shifts from less than 300 to 500 curves[4], as in Figure 2. This step was implemented to identify consumption patterns which should be treated as flat consumption cases (after the scaling, flat curves would be modeled as a vector with components equal to 1) and avoid them having an out of proportion impact on the clustering once that the normalization is applied and their shape magnified.

B. Clustering

1) Algorithms: We applied the most common clustering methods and rated them with the goal of finding clusters which group households based on their ability to single out peak consumption over the day. For this purpose, we examined different choices of parameters for the following clustering techniques : hierarchical clustering, K-Means and Self-Organizing Maps (SOM) dimension reduction followed by K-Means. We are aware of the curse of dimensionality, which has been covered by many authors such as more recently and particularly thoroughly by [17]. The latter implies that points in higher dimensionality cannot be differentiated as

[3] [16] determined that this value is on average 112 W in the case of households in New Zealand.

[4] 5.5% of all weekly curves

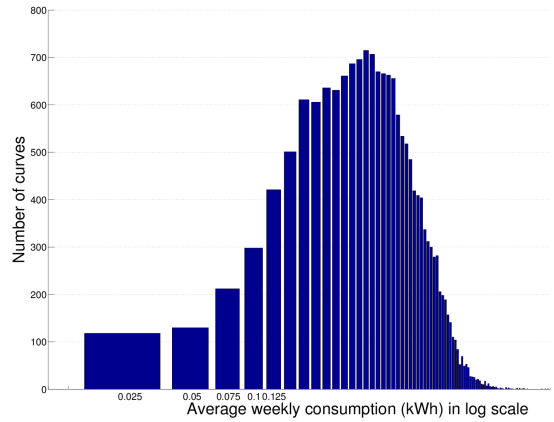

Fig. 2. Histogram of the distribution of the average weekly consumption using a log scale for 0.025 kWh bins.

summarized in Equation 1.

$$\lim_{d\to\infty} \mathrm{Var}\left(\frac{\|X_d\|}{\mathrm{E}\|X_d\|}\right) = 0 \Rightarrow \frac{D_{max} - D_{min}}{D_{min}} \to 0 \qquad (1)$$

The difference in performance of the whole clustering of the time series were evaluated against the extraction of a subset of features as listed in [18] or the usage of Pricincipal Components Analysis (PCA) to reduce the dimensionality of the data.

2) Number of clusters: We compared the performance of the clustering against the formation of 5 to 14 clusters, as more clusters would lead to over-fitting and overcome the purpose of simplifying the visualization of households consumption patterns. We also expected that a higher number of clusters would lead to some clusters containing very few load curves with isolated shapes, instead of being able to generalize and highlight common features of the data.

3) Combination of parameters: To decide upon the most appropriate clustering framework to suit our goal of identifying different peaks, we evaluated the following combinations of parameters:

- whole time series clustering and extraction of features
- Wiener filter window (no filtering, 2 to 5 samples windows length)
- number of clusters (from 5 to 14)
- combination of different clustering algorithms and distances as seen in Table I.

In particular, the whole time series clustering consisted of using 48-dimension load curves representing the average weekday consumption. We also needed to adapt the vectors when using the correlation and cosine distances. It required the data to be standardized column-wise, as the object we clustered were scaled between 0 and 1, implying that they had a relatively small standard deviation. We chose to evaluate the effectiveness of reducing the dimension of the input vectors to mitigate the curse of dimensionality. This was achieved by extracting 18 features, which comprehended statistical data over parts of the day (such as mean, standard deviation, min

TABLE I
COMBINATION OF THE EVALUATED CLUSTERING ALGORITHMS AND DISTANCE MEASURES

Clustering technique	Distance
SOM + K-Means	Manhattan
SOM + K-Means	Euclidean
K-Means	Manhattan
K-Means	Euclidean
K-Means	Correlation
K-Means	Cosine
Hierarchical	Manhattan
Hierarchical	Euclidean
Hierarchical	Correlation
Hierarchical	Cosine

and max) and ratios based on [18] and peak data, i.e., number of peaks during parts of the days. Alternatively, we selected the most significant PCA components (with contribution over 1%, which means 17 components).

After some preliminary testing, we adapted both the K-Means and hierarchical clustering methods in combination with the correlation and the cosine distances. Since the flat consumption patterns could not be singled out, we applied a two-phase clustering consisting of pre-applying the K-Means algorithm with the same settings but choosing the Euclidean distance to isolate the flat load curves. The clustering with the current choice of parameters was then carried out on the remaining curves.

The data were stored in a PostgreSQL database. Scripts to fetch and format the data were written in Python and Shell. For the clustering part, Matlab's implementation of the clustering algorithms was used, along with the SOM toolbox[5] and peakdet toolbox[6] for determining the peaks.

C. Similarity ranking

The usage of characteristic load profiles and their integration into an online portal can save the cost of re-clustering data for the new load curves. Selecting the cluster they belong to relates to selecting the most similar characteristic load profile, i.e., having the smallest distance. This allows to significantly reduce the cost of assigning a household's weekly consumption pattern to one of those clusters and thus permits a more scalable implementation. Also, it serves as a validation means for pondering the clustering accuracy.

For this purpose, we examined different distance measures such as the Manhattan, Euclidean and cosine distances along with the correlation between the household load curve and the load profiles. We focused on their ability to match a given load curve to the most similar reference curve.

IV. RESULTS ANALYSIS

A. Data

The work presented in [14] takes the seasonal component into consideration and more recently, the U.S. Energy Information Admnistration has reported that homes show seasonal variation in electricity use[7]. For this reason, 4 different subsets were built from the CER Irish dataset, which was collected from July 14, 2009 to December 31, 2010 as can be seen in Table II. DST dates for Ireland were used as benchmarks for separating winter from summer, i.e. October 25, 2009, March 28, 2010 and October 31, 2010.

TABLE II
WINTER AND SUMMER "TRAINING" AND "TEST"-SETS. THE TABLE ALSO CONTAINS THE TOTAL NUMBER OF DAILY LOAD CURVES AND THE CORRESPONDING REMOVED CURVES.

Start date	End date	# Days	# Weeks	Removed	Total
08/17/09	09/13/09	28	4	651	118271
08/17/09	10/31/10	287	41	720	118300
10/26/09	11/22/09	28	4	7153	1212138
10/26/09	12/31/10	215	31	7257	908177

B. Evaluation of the clustering

The selection of the most fitted clustering method and parameters relies on the target of identifying hurtful consumption behaviors as peaks. The performance of the clustering is in consequence based on how the peaks of different load curves match the peaks of the load profiles produced by the clustering. For this purpose, a binary vector l_i for the i^{th} load curve and a binary vector for its corresponding cluster c_i are built, taking the value 1 when a peak is at a given position. The ratio of matching peaks is computed as in Equation 2, where $< l_i, c_i >$ represents the inner product of the two binary vectors.

$$m_i = \begin{cases} \frac{<l_i, c_i>}{\sum_{k=1}^{48} l_i(k)} & \text{if } \sum_{k=1}^{48} l_i(k) > 0 \\ 1 & \text{if } \sum_{k=1}^{48} l_i(k) = 0 \text{ and } \sum_{k=1}^{48} c_i(k) = 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Then the "score" used is the average of Equation 2 over all curves in the considered dataset as in Equation 3. We refer to it as the peak match score.

$$\frac{1}{N} \sum_{i=1}^{N} m_i \tag{3}$$

We use a second score for rating the distinctiveness of the characteristic load profiles. This consists of summing the Hamming distances of all pairs of the binary representation of the cluster profiles.

The 20 top scoring configurations of parameters are highlighted in Table III. Although the scoring functions offer a quantitative way of evaluating the clustering, they merely provide a set of candidates that will be evaluated visually. The candidates for the best clustering parametrization combines the idea that the load curves have to match the characteristic load profiles and the latter have to be distinct from each other.

[5]http://www.cis.hut.fi/somtoolbox/
[6]http://www.billauer.co.il/peakdet.html
[7]http://www.eia.gov/todayinenergy/detail.cfm?id=10211#

TABLE III
20 TOP SCORING CONFIGURATIONS OF PARAMETERS FOR THE CLUSTERING

Type of clustering	Algorithm	Distance	Filt. Window	# Clusters	Peak Match Score	Distinctiveness Score
Whole clust.	K-Means	Correlation	5	14	0.2199	290
Whole clust.	K-Means	Correlation	5	13	0.21554	236
Whole clust.	K-Means	Correlation	4	14	0.21425	290
Whole clust.	K-Means	Correlation	5	12	0.21182	190
Whole clust.	K-Means	Correlation	4	13	0.20963	236
Whole clust.	K-Means	Correlation	5	11	0.2059	162
Whole clust.	K-Means	Correlation	4	12	0.20321	204
Whole clust.	SOM + K-Means	Euclidean	5	14	0.20179	273
Whole clust.	K-Means	Cosine	5	14	0.20156	260
Whole clust.	SOM + K-Means	Manhattan	5	14	0.19824	259
Whole clust.	K-Means	Correlation	4	11	0.19778	174
Whole clust.	K-Means	Euclidean	5	13	0.19673	192
Whole clust.	K-Means	Correlation	3	14	0.19624	290
Whole clust.	SOM + K-Means	Manhattan	5	13	0.19614	230
Whole clust.	K-Means	Cosine	5	13	0.19597	210
Whole clust.	K-Means	Correlation	2	14	0.19548	276
Whole clust.	K-Means	Cosine	5	12	0.1951	192
Whole clust.	SOM + K-Means	Euclidean	5	13	0.19502	230
Whole clust.	K-Means	Correlation	5	10	0.1946	136
Whole clust.	K-Means	Correlation	2	13	0.19417	238

Extracting features from the load curves leads to the issue of scaling them appropriately so that the components of the vector are not overpowering each other during the process of clustering, which is for example avoided when using the whole time series, as all readings are scaled. Overall, reducing the dimensionality from the 48-reading vector proved less successful as the scoring revealed that the peak match score was well below 10%. The presence of stacked versions of the same cluster was most prominent and hence, the distinctiveness of the clusters was not assured.

Also, distance measures such as the Euclidean and Manhattan distances tend to aggregate the points to the same cluster, as the notion of position of the peaks is absorbed through the summing. Thus, other attempts, such as transforming the load curve into a binary vector that marks the position of the peaks or padding the original load curve with its binary peak representation, did not succeed either.

Trading off these scores, K-Means with the correlation as a distance measure was selected. Also, the load curves were smoothed through the usage of a Wiener filter of window 3 (i.e. for each value of the load curve and 3 neighbors on the left on 3 on the right are used), which corresponds to using data in the scope of 1.5 hours around each measurement to correct the oscillations, which are considered as noise.

The "appropriate" number of clusters was 14 (15, if counting the group of flat curves that were excluded by the first clustering phase). The results can be seen in Figure 3, in contrast with Figure 4, where not all clusters are as distinct and we see stacked versions of the same flat cluster and overlapping peaks.

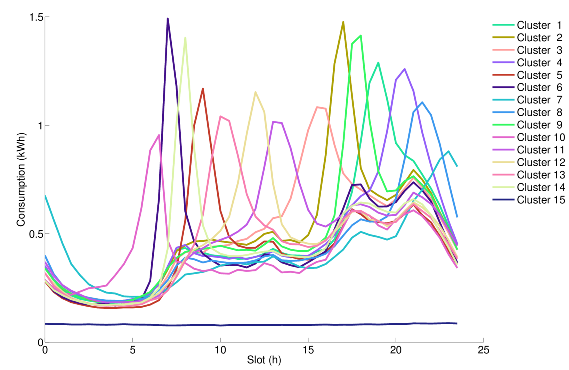

Fig. 3. 15 clusters (i.e., 14 + 1, obtained through first phase flat curves separation), K-Means, correlation, filter window = 3 on the training summer dataset. All characteristic load curves differ in the position of their peak.

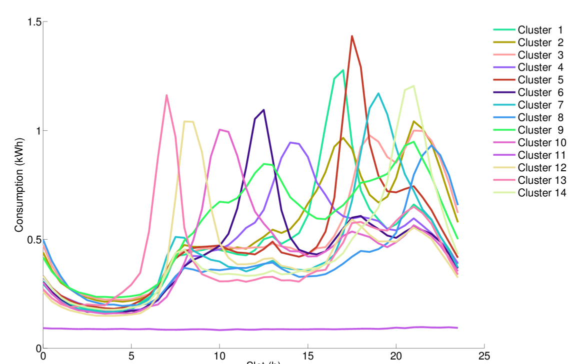

Fig. 4. 14 (i.e. 14 clusters with SOM + K-Means), Euclidean, filter window = 5 on the training summer dataset. Clusters 1 and 2 are not distinguishable as their peaks are located at the same positions.

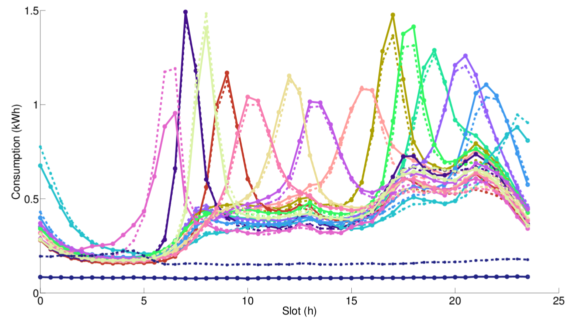

Fig. 5. Comparison between the clusters built from the "training" set from Figure 3. The dash-line curves represent the new cluster profiles using the summer "test" set using the cosine distance as the similarity measure.

C. Comparison with similarity distance ranking classification

Different distance measures to rank the similarity of the load curves to the cluster profiles were tested to determine the smallest distance to classify the curve into the right bin and assess the "quality" of the initial clustering on the training set. As can be seen in Figure 5, the best results are achieved with the cosine distance as the resulting cluster averages match the reference characteristic load profiles.

V. Conclusions and Outlook

We proposed a method to build cluster profiles with the objective of identifying hurtful behaviors from the utility companies' viewpoint. Once that the clusters are built, the classification of a household consumption pattern from one week to the other is achieved by ranking the similarity of each curve to the previously established "reference" consumption patterns. Overall, the clustering produces distinctive enough characteristic load profiles to target the discrimination of consumption patterns based on the peaks positions. "Classifying" households based on these profiles requires little overhead, which would permit an integration in an online portal and lead to more applications for the utility companies.

The segmentation of the households will allow the utility companies to get a better understanding of what consumption profiles exist among their customers and their proportion, instead of relying on an oversimplification of the consumer-base through the usage of the synthetic load profiles. Based on the energy provider's appreciation of what pattern is more hurtful, specific segments of customer can be easily selected and addressed. An application that could be foreseen would be to understand how the households' consumption evolves over time and target either the more stable households (i.e. selecting a threshold, as the percentage of weeks a specific household remains in the same cluster or simply select the top x number of households that have remained stable over time). This can be implemented in the frame of an awareness raising campaign as to maximize the chance of them reacting to a stimuli such as differentiated tariffs as a way of inciting load shifting.

We are aiming at expanding our analysis by verifying whether an underlying Markov chain could allow us to estimate the likelihood of a household to change their consumption from one week to the other, but also to highlight whether regional factors can impact the shape of the characteristic load profiles. Also, we are looking at finding out if household characteristics, which were also collected through surveys in the frame of the CER data collection can be mapped to the cluster profiles.

Acknowledgment

The authors would like to thank the Swiss Federal Office of Energy for their support and involvement in the Bits to Energy lab.

References

[1] M. Ghofrani, M. Hassanzadeh, M. Etezadi-Amoli, and M. Fadali, "Smart meter based short-term load forecasting for residential customers," in *Proc. NAPS*, Boston, Massachusetts, USA, 2011.
[2] F. McLoughlin, A. Duffy, and M. Conlong, "Evaluation of time series techniques to characterise domestic electricity demand," *Energy*, vol. 50, pp. 120 – 130, 2013.
[3] I. Sanchez, I. Espinos, L. Moreno Sarrion, A. Quijano Lopez, and I. Burgos, "Clients segmentation according to their domestic energy consumption by the use of self-organizing maps," in *Proc. EEM*, Leuven, Belgium, 2009.
[4] S. Ramos, V. Figueiredo, F. Rodrigues, R. Pinheiro, and Z. Vale, "Knowledge extraction from medium voltage load diagrams to support the definition of electrical tariffs."
[5] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Electric energy customer characterisation for developing dedicated market strategies," in *Proc. PowerTech*, Porto, Portugal, 2001.
[6] R. Chang and C. N. Lu, "Load profiling and its applications in power market," in *Proc. PES GM*, Toronto, Canada, 2003.
[7] S. Ramos and Z. Vale, "Data Mining techniques to support the classification of MV electricity customers," in *Proc. PES GM*, Pittsburgh, Pennsylvania, USA, 2008.
[8] F. Rodrigues, J. Duarte, V. Figueiredo, Z. Vale, and M. Cordeiro, "A comparative analysis of clustering algorithms applied to load profiling," in *Machine Learning and Data Mining in Pattern Recognition*, ser. Lecture Notes in Computer Science, P. Perner and A. Rosenfeld, Eds. Springer Berlin Heidelberg, 2003, vol. 2734, pp. 73–85.
[9] H. Liu, Z. Yao, T. Eklund, and B. Back, "Electricity consumption time series profiling: a data mining application in energy industry," in *Proc. ICDM*, Berlin, Germany, 2012.
[10] F. Iglesias and W. Kastner, "Analysis of similarity measures in times series clustering for the discovery of building energy patterns," *Energies*, vol. 6, no. 2, pp. 579–597, 2013.
[11] E. Keogh and S. Kasetty, "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration," in *Proc. SIGKDD*, New York, NY, USA, 2002.
[12] Commission for Energy Regulation, "Smart metering information paper 4," 2011.
[13] ——, "Electricity smart metering customer behaviour trials (cbt) findings report," 2011.
[14] A. Mutanen, "Customer classification and load profiling based on amr measurements," Department of Electrical Energy Engineering, Tampere University of Technology, Tech. Rep., 2010.
[15] G. Milligan and M. Cooper, "A study of standardization of variables in cluster analysis," *Journal of Classification*, vol. 5, no. 2, pp. 181–204, September 1988.
[16] M. Camilleri, N. Isaacs, and L. French, "The Baseload and Standby Power Consumption of New Zealand Houses," in *Proc. ACEEE*, Pacific Grove, California, USA, 2006.
[17] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality ?" in *Proc. SSDBM*, Heidelberg, Germany, 2010.
[18] C. Beckel, L. Sadamori, and S. Santini, "Towards Automatic Classification of Private Households Using Electricity Consumption Data," in *Proc. BuildSys*, Toronto, Canada, 2012.