Discovering Electricity Consumption Over Time for Residential Consumers Through Cluster Analysis

Tania Cerquitelli[§], Gianfranco Chicco[°], Evelina Di Corso[§], Francesco Ventura[§], Giuseppe Montesano*, Anita Del Pizzo*, Alicia Mateo González**, Eduardo Martin Sobrino**

[§]Dipartimento di Automatica e Informatica, Politecnico di Torino, Torino, Italy
{tania.cerquitelli, evelina.dicorso, francesco.ventura}@polito.it
[°]Dipartimento Energia "Galileo Ferraris", Politecnico di Torino, Torino, Italy
gianfranco.chicco@polito.it
*ENEL Foundation, Italy
{giuseppe.montesano, anita.delpizzo}@enel.com
**ENDESA Energia, Spain
{alicia.mateo, eduardo.martins}@enel.com

Abstract— **This paper presents an innovative and scalable methodology named CONDUCTS (CONsumption DUration Curve Time Series) to discover residential electricity consumption behaviours over time. CONDUCTS exploits data stream processing in time windows jointly with unsupervised machine learning on time-independent data. Specifically, time series consumption data for every consumer is split into N time windows. For a particular time window, a duration curve is calculated providing significant shape-based information disregarding temporal aspects. Each duration curve is sampled according to statistical characteristics and its relevant shape is captured. Therefore, every individual is represented by the evolution of N simplified duration curves. A cluster analysis, based on the K-means algorithm and the Euclidean distance, provides the different consumer profiles in a given time window. CONDUCT's current implementation runs on Apache Spark, a state-of-the-art distributed computing framework. As a case study, CONDUCTS has been experimentally assessed on the real hourly metered data collected in the time frame of one year for a large number of Spanish residential consumers. The experiments highlighted CONDUCT's ability to identify time-variable well-cohesive and well-separated groups of individual electricity consumption patterns with similar characteristics.**

Keywords— ***clustering; household; load pattern shape; time series; energy consumption; duration curve; smart metering***

I. INTRODUCTION

The knowledge of the consumers' electrical behaviour is a key aspect for various electrical system operators, such as distribution system operators, aggregators and retailers. One of the main interests of these operators is to obtain consistent groups of consumers exhibiting similar characteristics in the way they use electricity in a given time period. In the commercial dataset of an electric company, each consumer is associated with its contract information (i.e., voltage supply, contract power). Depending on the energy meter installed, information on the energy consumption in a given time period is also available. Grouping the consumers on the basis of contract information and energy consumption in a given time period is a simple way to proceed. However, in this way the grouping is found as a snapshot for the given period, and the details on the shape of the energy consumption during time are not considered.

This paper introduces a further way to use the time series data to provide significant shape-based information disregarding the temporal dimension. The proposed methodology, named CONDUCTS (CONsumption DUration Curve Time Series) exploits data stream processing jointly with unsupervised machine learning in order to identify patterns of individual electricity consumption and consumer behaviour over time. After the normalisation with respect to the contract power, the duration curve is constructed by ordering the normalised hourly consumption in the descending order. Statistics on deciles (features) are used to model the duration curve trend and used as inputs of the cluster analysis. The latter is performed over time windows considering weekdays and weekend days separately and for a specific number of weeks. Cluster analysis, based on the K-Means algorithm and Euclidean distance, is aimed at defining time-variable clusters representing the variety of the behaviour of the consumer groups at different time periods during the year. The CONDUCTS methodology exploits the computational advantages of state-of-the-art distributed computing frameworks: the current implementation runs on the widely-popular Apache Spark to quickly analyse very large data collections. As a case study, CONDUCTS has been validated on real hourly-metered electricity consumption collected in Spain. Experimental results, obtained on time series related to more than 500,000 consumers monitored every hour for one year, demonstrate the effectiveness of the proposed approach in discovering well-cohesive and well-separated groups of consumers with similar electricity consumption behaviours.

978-1-5386-1495-2/18/$31.00 ©2018 IEEE

The next sections of this paper are organised as follows. Section II discusses related works. Section III details the proposed methodology. Section IV reports and discusses the experimental results obtained on real energy consumption data for a large set of residential consumers. The last section contains the conclusions.

II. STATE-OF-THE-ART

The ongoing evolution of smart metering is enabling the operators to gather energy consumption data at shorter time steps. Thereby, the energy consumption *pattern* of each consumer can be defined with a given time step (e.g., hourly) in a given time interval (e.g., daily or weekly). If the consumption pattern of each consumer is regular during that time interval, namely, the hourly consumption in the different days (e.g., for the weekdays) has the same evolution, a typical weekday can be defined, and the consumer groups may be created by using clustering procedures [1]. This approach is usually applied to industrial and commercial consumers. The regularity of the patterns also enables the use of the Euclidean distance as the metric to determine the differences between pairs of patterns.

However, for household consumers the situation is different. Clustering the time series of household consumption is a very challenging task, because of many aspects concurring in making these time series very different even when the consumers have similar characteristics (family composition, number and size of the appliances). In fact, the evolution in time of the consumption depends on many behavioural aspects, among which the presence at home, the willingness to use certain appliances in a given day, the regular or irregular way to use the appliances at different hours, and the possible exploitation of timer-based commands for some appliances. In order to improve the quality of the analysis, the incorporation of exogenous variables could be useful. For example, in [2] socio-demographic factors such as the household size, net income and employment status have been considered to impact significantly on the electricity consumption. On another point of view, in [3] a supervised machine learning is used for revealing a number of household characteristics with satisfactory accuracy. In both cases, some information used in the characterisation or in the validation raises privacy concerns, so that the development of a consumer grouping approach purely based on the analysis of the time series of the energy consumption remains of real interest.

In a household, generally even the same consumer has no regular usage of the appliances in different days. Thereby, even the characterisation of an individual consumer is not straightforward. As such, trying and forecast the exact hourly consumption of an individual household on a multi-day time horizon may be a poorly formulated objective. For the same reason, the use of the Euclidean distance between two patterns would result in high distances between patterns with the same appliances just used at different times during the day. Some insights may be gained by identifying typical periods of the day related with common activities happening in a household. An example is given by the four time periods (overnight, breakfast, daytime, and evening) identified in [4] to represent different peak demand behaviour.

A better objective is to exploit the information embedded in the shape of the energy consumption without the strict reference to equidistant points in time. This can be done in different ways, but requires clarifying a basic assumption: the consumers are price takers, so that they do not change their behaviour depending on the electricity price. In this way, the electricity consumption time series can be used without adding details on the electricity prices and their forecasts. Moreover, in the present evolution of the consumer participation in demand response programmes [5], the consistent groups of consumers may be determined during normal periods without active demand response programmes. In this case, the electricity consumption time series are the ones contributing to form the baselines used to define the rewards following specific demand response actions [6].

One of the possibilities to overcome the regular sequence of the points in the time series is to use dynamic time warping (DTW), in which two time series are warped in a non-linear way by compressing and stretching the time axis to find the better match between each other [7]. However, dynamic time warping could exhibit scalability issues because of its computation time for large datasets (e.g., with millions of consumers) [8]. In addition, cases have been reported of DTW averaging inaccuracies in dealing with time series with k-means clustering [9]. However, the same issues have not been found by using k-medoids instead of k-means, and k-medoids DTW has been recently used for clustering household data in [10].

Another possibility to represent the time series information without the link to the time axis is to use probability-based data. In this case, the pattern data are transformed into probability densities by normalising them with respect to the total energy consumption in the time period of the pattern duration, then ordering the normalised data in the ascending order. This representation of the data is used in [11], where an adaptive K-means clustering algorithm is exploited to determine the number of clusters with a procedure starting from a set of cluster centres initialised with a standard K-means algorithm, and adding new cluster centres when a data violates a threshold based on the mean squared error with respect to the closest cluster centre.

A parallel research effort has been carried out to design and develop innovative systems based on Big Data technologies to provide different analytics services. Distributed and parallel approaches have been proposed in recent years, including widespread Big Data frameworks like Apache Hadoop [12], which provides the most popular MapReduce implementation, and its many extensions and related projects, such as Apache Spark [13]. However, the exploitation of such distributed frameworks in the energy domain is challenging, because it requires a high level of expertise in computer science to address both technical issues and the last cutting-edge technologies in a properly way. Some research effort solutions

have been devoted to designing a general purpose engine [14] or tailored to a given application domain, such as thermal energy consumption [15] [16] and residential energy use [17].

III. THE CONDUCTS METHODOLOGY

The components of the CONDUCTS methodology, as well as the interactions between such components, are shown in Fig. 1. These components perform two main tasks: (a) data stream preprocessing, and (b) self-tuning individual profile characterisation. The current implementation of CONDUCTS runs on the Apache Spark framework, supporting parallel and scalable processing for analytics activities.

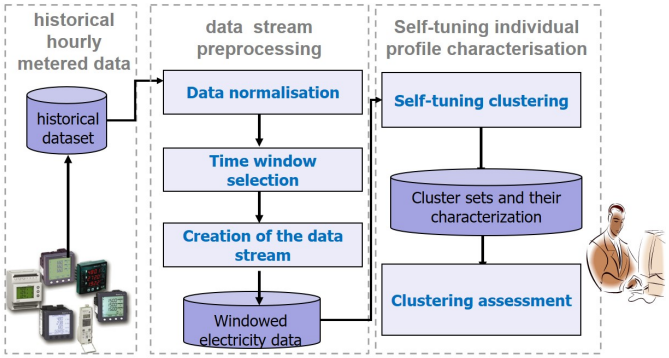

Fig. 1. The CONDUCTS Architecture.

A. Data stream preprocessing

This component addresses three tasks that have been proved to be crucial in real-world sensor-provided energy data: (i) *data normalisation*, (ii) *time window selection*, and (iii) *creation of the data stream*.

Data normalisation. The choice of the type of data normalisation is a key point, as the results are strongly affected by the adopted solution. This paper is focused on discovering shape-based patterns, removing the effect of the consumption level. For this reason, CONDUCTS integrates the *normalisation with respect to the contract power*. Additionally, it allows to easily identifying outliers (as consumption greater than one after the normalisation) by preserving the data distribution.

Time window selection. As the smart meter records the consumption on an hourly basis, it is needed to deal with a huge amount of data. The CONDUCTS engine has been designed to perform the data analytics task through the data stream analysis over a time window. Conceptually, the selection of the time window parameter is a strategic decision and its optimisation is out of the scope of this paper. Nevertheless, the time window length should be able to capture different regimes of the external variables as temperature or weather conditions and has to be long enough in order to include an adequate number of points for clustering purposes. Furthermore, weekdays and non-common days (i.e. weekend days and a number of special days, corresponding to holidays or pre-holidays) have been analysed separately.

Creation of the data stream. This step identifies the set of features to be used as input data for the cluster analysis. Specifically, for each consumer the hourly data belonging to the same time window have been ordered in descending values for a sample of 10% of the consumers. For these duration curves within a given time window, the average variations between each value and the following one have been calculated. Then, these variations have been summed up sequentially (obtaining a monotonically increasing curve). The resulting curve has been transformed into a curve with values ranging from 0 to 1 on the vertical axis, by dividing each value by the last value. This curve has been interpreted as a cumulative distribution function (CDF), from which the deciles have been identified. The cut points on the horizontal axis have been identified at the end of the last 9 deciles (excluding the first decile, having low variations). Finally, the 9 cut points have been applied to the duration curve of the normalised hourly energy consumption of each consumer, and the 9 selected features representing each consumer in the given time window have been determined as the average values of the duration curves inside the last 9 deciles.

B. Self-tuning individual profile characterisation

This CONDUCTS activity entails the discovering of groups of individual consumption profiles with a similar trend in a specific time window providing the self-tuning of the desired number of clusters. It includes three main components: (i) *a self-tuning clustering algorithm*, (ii) *clustering characterisation*, and (iii) *clustering assessment*.

The *self-tuning clustering* algorithm integrates a partition clustering algorithm and a strategy to automatically discover the desired number of clusters. In particular, CONDUCTS uses the K-means algorithm [18] with Euclidean distance, which is the most popular clustering algorithm in the literature. The objective of this step is to identify groups of consumers with similar duration curve profiles in a particular time window.

The K-means algorithm splits the input dataset into K groups, where K is a parameter that must be adjusted in advance. Each group is represented by its centroid computed as the average of all sampled consumer's duration curves in the cluster. The selection of the parameter K is crucial due to the fact that an automatic strategy has to be applied in the CONDUCTS methodology. In this sense, CONDUCTS provides a self-tuning strategy. Specifically, for a given time window, the clustering session is performed for a wide range of K values, calculating for every value of K its SSE (Sum of Squared Errors) associated. The obtained SSE values are plotted against K to reproduce the graph on which the well-known "elbow" (or "knee") criterion [19] is applied: the optimal value of K is selected where the gain from adding a cluster becomes relatively low, or in other words the SSE reduction is considered to be not worth enough compared with the increase of complexity in the clustering.

For what concerns *clustering characterisation*, CONDUCTS characterises the cluster set through different

methods to highlight the quality of the identified partition (e.g., the ability of determining well-separated and cohesive groups of individual consumption profiles), as well as to provide interesting insights on the nature of the individual consumption profiles.

Specifically, CONDUCTS provides:

- *Centroids-based characteristics* to graphically show an overview of the discovered cluster sets. Specifically, the duration curves corresponding to each cluster centroids are plotted to provide a quick and easy visualisation of the different average individual profiles representing the identified groups.
- The *boxplot distribution* [20] for duration curves and the corresponding daily time series. CONDUCTS exploits the boxplot (also known as whiskers plot) to graphically show groups of numerical data (duration curves as well as the corresponding time series) through their quartiles.
- *Scatter plot of the duration curves* to represent the relations between the maximum hourly consumption and the average consumption, for each cluster.

The *clustering assessment* component evaluates the ability of the CONDUCTS engine to correctly perform the cluster analysis of individual consumers by analysing data in a given time series. To this aim, CONDUCTS integrates the Silhouette index [21] to evaluate the quality of the cluster models, which measures both intra-cluster cohesion and inter-cluster separation. The objective is to evaluate the appropriateness of the assignment of a consumer's duration curve to a cluster rather than another one. This index evaluates the quality of each individual separately. The Silhouette coefficients take values in [-1,1]. Negative and positive Silhouette values represent wrong and good duration curve placements, respectively. Hence, the ideal clustering algorithm splits the data in a set of clusters such that all clusters have a Silhouette value equal to 1. As an order of magnitude, average Silhouette values around 0.2 are already considered good values representing clustering results [22].

IV. Experimental Results

Wide ranges of experiments were performed in order to assess the effectiveness of CONDUCTS in discovering groups of consumers with similar electricity consumption behaviour. The experiments have been carried out on real hourly-metered data collected during one year (from 2016-05-01 to 2017-04-30) for 565,662 Spanish residential consumers. The input data include the contract power used to normalise the data and other information out of the scope of this paper.

The current implementation of CONDUCTS is a project developed in Scala exploiting the K-Means algorithm available in MLlib. Experiments have been performed on a 3.6GHz quadcore Intel Core i7 PC with 32Gbyte main memory running a standalone Apache Spark 2.1.0.

CONDUCTS is configured with different time windows in order to separate working and non-working days. A two-weeks time window is selected for the working days. It is important to highlight that ten days (from Monday to Friday) are included to be clustered in most of the time windows. However, the presence of special days as bank holidays (e.g., Christmas) can modify the general behaviour of these days, so a particular filtering is carried out in order to eliminate them. Additionally, the time windows in the summer period requires a special pre-processing as the behaviour in this period is totally different to the rest of the year. In particular, a constraint is included in order to avoid that working days in August could share time windows with working days from adjacent months. Non-working days are grouped together with the special days, and a single time window is defined for every month.

The results of CONDUCTS over the data set period are 38 time windows: 26 defined by working days, and 12 by non-working days. To set automatically the desired number of clusters for the K-means algorithm, the method explained in Section III.B is exploited. Fig. 2 displays the evolution of the SSE over different values of K for the time window (2,3). After running this method through the 38 time windows, no significant differences appears between them. In order to simplify the methodology, a unique K value is finally adopted for all time windows. The candidate for K chosen for this study has been set to K = 6, that is, the value found for the large majority of the time windows with the application of the "elbow" criterion. The latter especially does not see any improvement when an extra cluster is added and has therefore been chosen for this test.

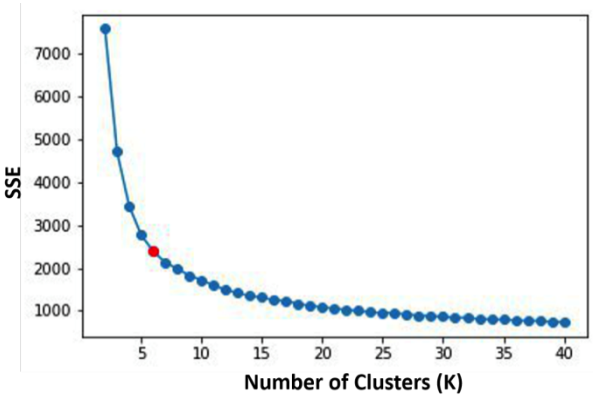

Fig. 2. SSE trend analysis.

A. Clustering assessment and characterisation

The objective of this section is to analyse the quality of the cluster analysis yielded through CONDUCTS. A first analysis based on the Silhouette trend over all time windows has been performed for a random sample of 10,000 consumers.

Fig. 3 shows a stable evolution of the silhouette values through time windows for the working days. The results show consistently similar values of the indicator throughout the time windows used in the analysis. These results demonstrate the ability of the CONDUCTS methodology in discovering a good set of groups of duration curves. Based on the Silhouette

values, all the discovered partitions include cohesive and well-separated groups of duration curves.

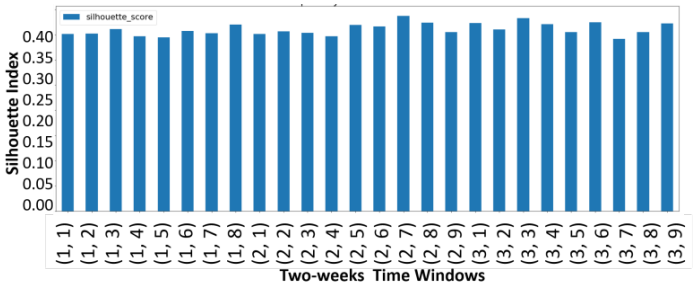

Fig. 3. Silhouette values for clusters in each working day time window.

CONDUCTS provides different insights to characterise the cluster set discovered over each time window. Here we discuss the results obtained on time window (2,3) corresponding to the first two weeks of June. Table I summarises the number of consumers grouped in each cluster, while Fig. 4 shows the cluster centroids represented by the 9 features on the horizontal axis. CONDUCTS discovers three large clusters, two medium clusters and a small one. Indicatively, the large clusters include consumers with low energy *utilisation factor* (the energy consumption to contract power ratio, based on the cluster centroids shown in Fig. 4) while medium clusters correspond to consumers with medium energy profile, and the smallest cluster represents the subset of consumers with high utilisation factor profiles. However, it has to be noticed that the partitioning is obtained by using shape-based information, not the energy utilisation factor directly.

TABLE I. NUMBER OF CONSUMERS FOR EACH CLUSTER

Cluster ID	Size
0	108,993
1	36,489
2	174,270
3	90,207
4	147,856
5	7,847
Number of consumers	**565,662**

Fig. 5 shows the boxplot of the duration curve of the 9 selected features driving the cluster analysis. The clustering results identify well-separated clusters. Fig. 6 replicates the cluster composition for the normalised daily energy consumption and reports a boxplot for each cluster. The clustering results provide a clear separation among the clusters also with respect to this variable.

Fig. 7 shows the scatter plot representing the relations between the maximum and the average normalised hourly energy consumption for each cluster. All entries belonging to the same cluster are represented with the same colour. Furthermore, the different colour intensity highlights the presence of possible outliers (from the plot with $a = 0.2$) or a view with more transparency (with $a = 0.002$) that indicates the relatively small overlap between the points belonging to the different clusters. It can be noted that the 9-dimensional features do not contain explicitly neither the average nor the maximum normalised hourly energy consumption.

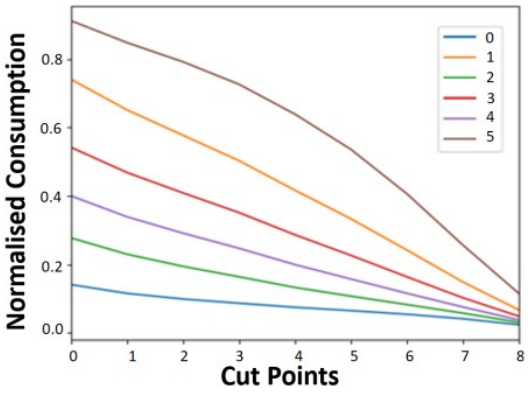

Fig. 4. Centroid distribution for each cluster.

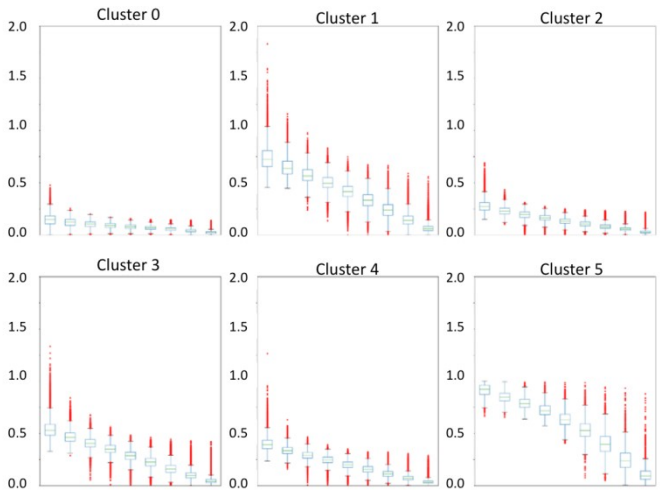

Fig. 5. Boxplot of the nine features for each cluster.

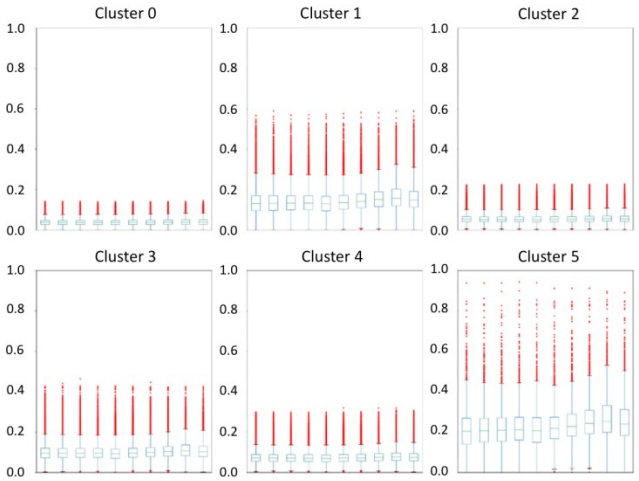

Fig. 6. Normalised daily energy consumption boxplot for each cluster.

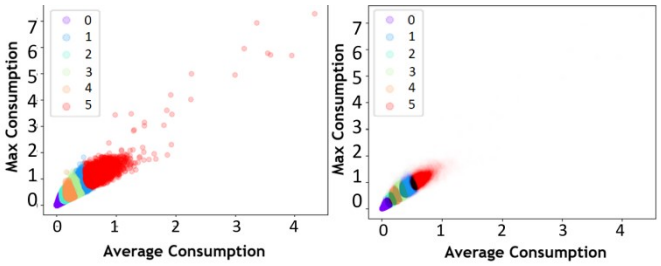

Fig. 7. Scatter plot for the average and maximum normalised hourly energy consumption of each consumer: colour intensity a = 0.2 (left) vs. colour intensity a = 0.002 (right).

V. Conclusion and Future Reasearch Directions

This paper proposed an innovative scalable methodology, named CONDUCTS, to discover electricity patterns over time. CONDUCTS has been tested on a large volume of hourly electricity consumption related to more than 500,000 Spanish residential consumers monitored for one year. The results demonstrated the capability of CONDUCTS to discover good partitions over time of consumers with a similar behaviour expressed by the duration curves of their hourly consumption. A possible extension of this work is the design of a methodology to build the electricity profile storyboard for each individual based on the cluster results obtained on consecutive time windows. The methodology can be used in other kind of studies, modifying the data normalisation and/or the time windows. Depending on the requirements to be fulfilled, these parameters could be adapted in order to split the data into smaller slices to be analysed in further steps.

References

[1] Chicco, G., Napoli, R., & Piglione, F. (2006). "Comparisons among clustering techniques for electricity customer classification". IEEE Transactions on Power Systems, 21(2), 933-940.

[2] Hayn, M., Bertsch, V., & Fichtner, W. (2014). "Electricity load profiles in Europe: The importance of household segmentation". Energy Research & Social Science, 3, 30-45.

[3] Beckel, C., Sadamori, L., Staake, T., & Santini, S. (2014). "Revealing household characteristics from smart meter data". Energy, 78, 397-410.

[4] Haben, S., Singleton, C., & Grindrod, P. (2016). "Analysis and clustering of residential customers energy behavioral demand using smart meter data". IEEE Transactions on Smart Grid, 7(1), 136-144.

[5] Albadi, M. H., & El-Saadany, E. F. (2008). "A summary of demand response in electricity markets". Electric Power Systems Research, 78(11), 1989-1996.

[6] Wijaya, T. K., Vasirani, M., & Aberer, K. (2014). "When bias matters: An economic assessment of demand response baselines for residential customers". IEEE Transactions on Smart Grid, 5(4), 1755-1763.

[7] Berndt, D. J., & Clifford, J. (1994, July). "Using dynamic time warping to find patterns in time series". In KDD workshop (Vol. 10, No. 16, pp. 359-370).

[8] Bartoš, T., & Skopal, T. (2012, August). "Revisiting techniques for lower bounding the dynamic time warping distance". In International Conference on Similarity Search and Applications (pp. 192-208). Springer, Berlin, Heidelberg.

[9] Niennattrakul, V., & Ratanamahatana, C. A. (2007, April). "On clustering multimedia time series data using k-means and dynamic time warping". In Multimedia and Ubiquitous Engineering, 2007. MUE'07. International Conference on (pp. 733-738). IEEE.

[10] Teeraratkul, T., O'Neill, D., & Lall, S. (2017). "Shape-Based Approach to Household Electric Load Curve Clustering and Prediction". IEEE Transactions on Smart Grid, in press, doi:10.1109/TSG.2017.2683461.

[11] Kwac, J., Flora, J., & Rajagopal, R. (2014). "Household energy consumption segmentation using hourly data". IEEE Transactions on Smart Grid, 5(1), 420-430.

[12] Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. Hadoop Project 11:21.

[13] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., et al. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. p. 2-2.

[14] Zulkernine, F. H., Martin, P., Zou, Y., Bauer, M., Gwadry-Sridhar, F., & Aboulnaga, A. (2013, June-July). "Towards cloud-based analytics-as-a-service (CLAAAS) for big data analytics in the cloud," Proc. IEEE International Congress on Big Data, BigData Congress, pp. 62–69.

[15] Anjos, D., Carreira, P., & Francisco, A. P. (2014, June-July). "Real-time integration of building energy data," Proc. IEEE International Congress on Big Data, Anchorage, AK, pp. 250–257.

[16] Acquaviva, A., Apiletti, D., Attanasio, A., Baralis, E., Bottaccioli, L., Castagnetti, F. B., Cerquitelli, T., Chiusano, S., Macii, E., Martellacci, D., & Patti, E. (2015, June). "Energy signature analysis: Knowledge at your fingertips". IEEE International Congress on Big Data (BigData Congress), pp. 543-550.

[17] Wang, C. de Groot, M., & Marendy, P. (2009, July). "A service-oriented system for optimizing residential energy use," Proc. IEEE International Conference on Web Services (ICWS), Los Angeles, CA, pp. 735–742.

[18] Juang, B. H., & Rabiner, L. R. (1990). "The segmental K-means algorithm for estimating parameters of hidden Markov models". IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(9), 1639-1641.

[19] Tan, P. N., Steinbach, M., & Kumar, V. (2005). "Introduction to Data Mining". Addison-Wesley, 2006.

[20] S. M. Ross (2000). "Introduction to probability and statistics for engineers and scientists" (2. ed.). Academic Press, 2000.

[21] Rousseeuw, P. J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". Journal of Computational and Applied Mathematics, 20, 53-65.

[22] Di Corso, E., Cerquitelli, T., & Ventura, F. (2017, April). "Self-tuning techniques for large scale cluster analysis on textual data collections". In Proceedings of the ACM Symposium on Applied Computing (pp. 771-776).