Clustering-Based Assessment of Residential Consumers from Hourly-Metered Data

Tania Cerquitelli[§], Gianfranco Chicco[°], Evelina Di Corso[§], Francesco Ventura[§], Giuseppe Montesano*, Mirko Armiento*, Alicia Mateo González**, Andrea Veiga Santiago**

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

Abstract— **This paper addresses the methodology for determining suitable groups of residential consumers, based on time series of their hourly energy consumption and contractual data. Salient aspects are the discussion on the importance of the data representation in terms of data normalisation, choice of the appropriate features to be used as inputs in clustering procedures, and computation of clustering validity indicators. The analysis is carried out on real hourly-metered electricity consumption data of 10,000 residential consumers. We discuss the main insights obtained with the application of conventional approaches based on time series data handled with different distance metrics (e.g., Euclidean distance and dynamic time warping) and alternative approaches exploring data transformations, among which the CONsumption DUration Curve Time Series (CONDUCTS) methodology proposed by the authors.**

Keywords—clustering; household; load pattern shape; time series; energy consumption; duration curve; dynamic time warping

I. Introduction

The growing availability of data gathered from smart meters requires appropriate procedures for extracting useful information from a continuous flow of metered data. Clustering techniques can be used for the analysis of a large amount of data (e.g., coming from the nation-wide installed meters) [1]. However, a key aspect of the analysis is the definition of the features to be used as inputs to the clustering procedure. This definition is not unique and depends on the type of consumers analysed (e.g., residential, industrial), the time step of the data available (typically 15, 30 or 60 minutes), and other information that may be available from the company's databases.

Among the various types of consumers, individual residential consumers are the most challenging ones to be addressed, as their consumption depends on a number of behavioural aspects conditioning the regular or irregular way to use the appliances during time, as well as other factors such as number of inhabitants, net income, age of the persons in the family, employment status, and other socio-demographic information [2]. Further situations are emerging in which the household consumption has to be handled together with local generation (e.g., rooftop photovoltaic systems) [3], making the net power consumption (that is, demand minus local generation) quite different with respect to the past. In [4], five household segmentation strategies are developed based on an encoding system based on a load shape dictionary containing most frequent usage patterns. In [5], a finite mixture model-based clustering is used to deal with both continuous and categorical data.

The variability of the energy consumption patterns for residential consumers requires specific formulations of the input data for clustering, as classical assumptions such as the use of time series together with the Euclidean distance metric are rather inappropriate. In fact, for example the use of the Euclidean distance leads to high distances if two patterns contain a base load and a similar peak located in different time periods in the two patterns. However, this *diversity* among the positions of the peaks could be only due to the non-regular usage of the same appliance during the day and appears even for two patterns of the same consumer. To avoid the effect of such diversity on the results of load pattern grouping, it is possible to use specific metrics such as dynamic time warping [6], which tends to create an optimal alignment between the patterns by stretching the horizontal scale.

This paper presents a comparison among different ways to represent the data of residential consumers for creating consistent consumer groups through cluster analysis. The available data include the time series of household consumption and their contractual power. Neither categorical data nor socio-demographic information is available. As a real case study, is it illustrated how to effectively analyse the contractual data and the time series of the hourly energy consumption gathered for 10,000 consumers for one year. The importance of data normalisation is discussed by providing specific examples. The effectiveness of using shape-based representations constructed by using the time series data (from conventional regular patterns to the application of dynamic time warping) and by applying the methodology named CONDUCTS (CONsumption DUration Curve Time Series) [7] developed by the authors is assessed and compared. In addition to data normalisation issues, the contents presented

978-1-5386-5326-5/18/$31.00 ©2018 IEEE

include, the definition of the type and number of features to be used in the analysis, the choice of the number of clusters, the execution of the clustering procedures, the evaluation of clustering validity indicators that express the quality of the clustering results.

The next sections of the paper are organised as follows. Section II recalls the data representation aspects, with reference to the data normalisation used in this paper. Section III introduces the different features used in this paper for comparative analysis, as well as the distance metrics and the clustering validity indicators used to quantify the effectiveness of the clustering results. Section IV shows the results and the related discussion. Section V contains the conclusions.

II. Data Representations for Residential Consumers

Data normalisation is applied to the input data. Normalisation is a significant aspect for dealing with time series of the electrical demand, as the normalised time series may look different and as such the consumers may be grouped in different ways by the clustering algorithms. Two normalisation techniques are considered in this paper:

- P_{ref}: defines a *reference value* as a user-defined scaling factor, e.g., the *contract power*, and divides the time series values by the reference value. This solution creates scaled time series, for which the relevant aspect is the *shape* of the electrical demand – a more qualitative and time-dependent aspect of the nature of the energy consumption, regardless of the amplitude of the time series, whose information is carried by the contract power [8].
- Max-min binormalisation: the time series is mapped into the interval [0,1] on the vertical axis, through a linear transformation that associates the value 0 to the lowest value, and the value 1 to the highest value. This solution creates a stretching of the time series, losing the uniformity of the values on the vertical axis for different time series and changing the relations among the shapes of the time series.

III. Configuring The Cluster Analysis: Selection Of Features, Distance Metrics and Validity Indices

A. Time windows-based approach

The residential consumers may have different consumption patterns depending on the period of the year (e.g., winter, summer), in which some appliances used could be different, and the lifestyle of the consumers could change. As such, mixing the data for the whole year is not effective. Furthermore, some differences could arise between weekdays and weekend periods. On these bases, a suitable way is to identify sub-periods (e.g., months or other sub-partitioning). This paper exploits the concept of creating time windows of user-defined duration, in which the consumption patterns belonging to the same time window are handled together. This approach requires the definition of the time window parameter w_s, which determines the temporal context of interest for the analysis. The time windows can be of different length depending on the period analysed. More specifically, if the time window is very short, only the most recent consumption of the customer will be analysed, but similar behaviours could appear in adjacent time windows and many similar cases could be generated in the study. Instead, a too large time window allows analysing many data on past customer electricity consumption, but it may introduce noisy information in the cluster analysis. In this paper, the value for w_s has been set to two weeks for the weekdays and to one month for the weekend days. Furthermore, the presence of special days such as religious holidays at fixed dates, bank holidays, and local festivities, has been considered by including these special days in the category of weekend days.

Every time electricity consumption is collected, one time window over the data stream is considered for the cluster analysis task. This time window contains a snapshot of the electricity consumption monitored in all the hours belonging to the time window. It describes the recent past electricity consumption of the consumer, and consequently, characterises its recent consumption pattern.

For each time window, the input data is composed of a matrix with one consumer on each row and its consumption data on the columns (introduced as hourly time series, or as different features deriving from the data feature selection techniques executed in the pre-processing phase), as described below.

B. Feature selection

Two types of features are considered in this paper:

1. The time series of the *hourly* data generated by the electricity meters (stream analysis).
2. A selected set of *features* originated by the normalised duration curves constructed by using the hourly data metered in the same time period. The rationale of using this type of features is of specific relevance for the residential consumption patterns. In fact, in the type of analysis considered in this paper, the residential consumers that use the same appliances in different periods of the day can be considered as similar. This happens because the focus of this paper is to categorise the consumers from the whole pattern and not to address specific timing issues relevant to the formulation of demand response programmes [9]. The generation of the selected features has been described in [7] and is briefly recalled here. Let us consider a two-week period with 240 hourly data from weekdays. These data are first ordered in the descending order to construct the duration curve. Then the variations determined from two successive data on the duration curve are considered for each consumer, and the cumulative distribution function (CDF) of the average variations for all consumers is calculated at each point of the duration curve. The resulting CDF has 239 points (because of the calculation of the variations, the first point is excluded). This CDF is then cut into a given number of proportional intervals (e.g., deciles) on the vertical axis, and the corresponding *cut points* are selected. In this paper, *nine* cut points have been selected, by excluding the first and the last point of the deciles limits. A key aspect is that the definition of the cut points is done by considering the whole set of consumers in the time window. After that, the same cut points

are used to pick up the nine values referring to each consumer. These nine points are the selected features that are assumed to describe the behaviour of each consumer.

C. Distance metrics

The particularity of the residential consumers with respect to other types of consumers described above also reflects on the type of distance that can be used to establish the similarity among the patterns or the sets of selected features. In particular, considering the time domain data, the use of the Euclidean distance would generate high distances whenever even the same appliances are used in different time periods. In this case, better approaches based on dynamic time warping (dependent on the rescaling of the time axis) have been proposed, as mentioned in the Introduction.

However, if the nine selected features are extracted as indicated above, for each consumer these features are by definition monotonically decreasing. In this case, the use of the Euclidean distance can be reasonable. In addition, it could be expected that the use of the dynamic time warping does not provide results so different with respect to the use of the Euclidean distance, as there are no peaks requesting significant distortion of the horizontal axis. Thereby, to support this conjecture, in this paper some comparisons are shown between cases using the Euclidean distance and the dynamic time warping. The use of dynamic time warping requires the specification of a parameter, here called d_z, where z is the maximum "distance" among the horizontal points at which the non-linear distortion of the horizontal axis is enabled. In practice, d_1 means no distortion, and the corresponding alignment between pairs of patterns is the same as the one used to calculate the Euclidean distance. The value z is user-defined, depending on the nature of the data.

The organisation of the tests is described below. For different combinations of normalisation and alternative, the K-Means clustering is executed. The output of the clustering algorithm is a column vector which contains, for each consumer, the number of the cluster to which the consumer has been assigned. Finally, the clustering validity indicators are calculated for each solution, by using different types of data input. In this way, it is possible to identify the most effective combination of normalisation and alternative.

D. Clustering validity indicators

Once the K clusters have been formed starting from the collection of load patterns, the clustering outcomes are subject to clustering validity assessment, by using two indicators based on the calculation of the widely used *silhouettes* [10]. The silhouette index is a quality measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring cluster. Two indicators based on the silhouettes are used in this paper:

- The average silhouette index (*ASI*), calculated by averaging the silhouettes over the entire cluster set.
- The global silhouette index (GSI), which considers the possible imbalance number of elements in each cluster. Clusters with large number of load patterns are penalized in the *GSI* computation. For both indicators, higher values correspond to better clustering validity.

IV. Case Study: Cluster Characterisation and Validity Assessment

Many experiments have been performed on a real dataset to discuss three main issues: (i) cluster configuration, (ii) assessment of the cluster sets, and (iii) cluster characterisation. The experiments have been carried out on real hourly-metered data collected for one year (from 2016-05-01 to 2017-04-30) on 10,000 Spanish residential consumers. The input data include the contract power used to normalise the data. The entire year has been partitioned into 38 time windows, of which 26 time windows are defined with the hourly data of two weeks of working days each, and 12 time windows contain the hourly data of one month of non-working days. This section presents the results for a representative time window in the first two weeks of June.

The current implementation of CONDUCTS is a project developed in Scala exploiting the K-Means algorithm available in MLlib [11]. 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. To perform the set of experiments with dtw we used the TSclust package [12] available in R [13]. The TSclust package includes a wide set of dissimilarity measures tailored to time series. Within TSclust, we exploited the R package dtw [14] to run the experiments named dtw_d_z.

A. Clustering algorithm and choice of the number of clusters

The configurations analysed in this paper mainly focus on the normalisation and distances, with the use of one clustering method that is able to provide relatively uniform partitions of the datasets, rather than isolating specific outliers. The K-means algorithm [15] has been used to create the consumer groups. Comparisons with other clustering algorithms are left to a follow-up of this contribution. K-means is one of the most popular clustering algorithms capable to identify the cluster set in a limited computational time by producing quite good results in many domains. One of the biggest drawbacks of K-means is that it requires the number of clusters K to be a-priori specified. To address this issue, the elbow criterion is exploited by analysing the Sum of Squared Errors (SSE) [15] trend against K. Specifically, the SSE index measures the cluster quality in terms of cluster cohesion. The total sum of squared errors is computed for all consumers in the dataset, where for each consumer (time series) the error is computed as the squared Euclidean distance from the closest centroid. The SSE tends to decrease by increasing K because smaller clusters are discovered which are naturally more cohesive. Here, we selected as possible good values of K the coordinates where the marginal decrease in the SSE curve is maximized – the *elbow* zone. Fig. 1 shows the SSE trend for K in the range [2-40] computed on the time window under analysis. The elbow criterion suggests two possible values for the desired numbers of clusters (K = 6 and K = 7). Nevertheless, after running this method through the 38 time windows, no significant difference appears between them in almost all the cases. Thus, $K = 6$ and $K = 7$ have been used in the tests, also for the sake of comparison among the corresponding results.

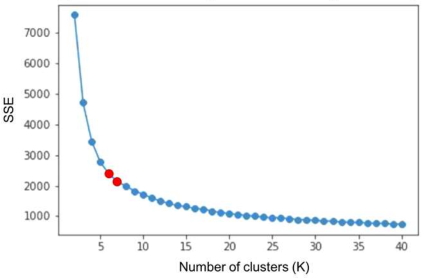


Fig. 1. Definition of the number of clusters by using the elbow criterion.

B. Clustering results and discussion

The K-means clustering has been executed by using the data inputs indicated by the configurations shown in Table I. In the time window under analysis, the number of features is 240 for the hourly data and 9 for the cut points identified as described in Section III.B. The dynamic time warping has been used with window size values $d_2 = 2$ and $d_3 = 3$ in the case with 9 cut points, and with the window size value $d_{10} = 10$ for the cases with 240 features. Table I shows the clustering validity calculated by using the *ASI* and *GSI* indicators. The values highlighted in bold are further discussed with specific plots.

Let us start with the results referring to the Euclidean distance. Configuration C5, for $K = 6$, exhibits better indicators than configuration C6 with $K = 7$, and is taken as the reference for the dataset with 9 cut points and contract power used as the normalisation factor. In the same way, Configuration C7 is better than Configuration C8 and is taken to represent the max-min normalisation. Fig. 2 and Fig. 3 show the silhouette plots with the corresponding distribution for each cluster for configurations C5 and C7, respectively. Configuration C5 appears to be overall better than configuration C7, although all clusters show a prominent silhouette with only a few consumers with silhouette values lower than zero. As shown in Fig. 2b the medians of the silhouette values for each cluster in C5 assume values higher than the corresponding ones in C7. Moreover, configuration C7 is biased by the presence of an anomalous and relatively small cluster (Cluster #3) with silhouette values mostly saturated at a very high value.

Furthermore, configuration H2 is better than configuration H1 to represent the case with 240 hourly values and contract power as the normalisation factor. In addition, compared to all results reported in Table I, configuration H2 has the highest value of *ASI,* and could be considered as the best solution. However, the associated *GSI* value is very low (close to zero). This situation suggests a dedicated analysis of the silhouette plots and the corresponding distributions, whose results are reported in Fig. 4. The silhouette values in Configuration H2 are largely related to the presence of a very large dominating cluster (Cluster #0), which collects more than 80% of consumers, while the other six clusters have very poor silhouettes. In fact, the *GSI* measure is very low, since it penalises cluster with unbalanced numbers of consumers. Thereby, Configuration C5 can partition the consumers into balanced and well-separated clusters with respect to Configuration H2.

About the solutions obtained with the dynamic time warping, by using the dataset with 9 cut points and contract power as the normalisation factor, the solutions obtained with time window values $d_2 = 2$ (or $d_3 = 3$) are slightly worse than the Configuration C5 in terms of the indicator *ASI*, and relatively worse in terms of the indicator *GSI*. This can be explained by the nature of the specific dataset, in which all the values corresponding to the 9 cut points are monotonically decreasing by construction, so that there is little room to improve the situation through warping. On the other hand, when the 240 hourly values are considered with the contract power normalisation and time window $d_{10} = 10$, both indicators *ASI* and *GSI* become extremely poor. The time series are typically composed of many local peaks and the application of the dynamic warping process to a dataset of the considered size has proved to be rather ineffective. With other values of size window used for the dynamic time warping the situation in general does not improve, as shown in Fig. 5. The cluster set cardinality is quite balanced, however the silhouette value distribution presents very low median values, and the interquartile range is lower than zero in each cluster.

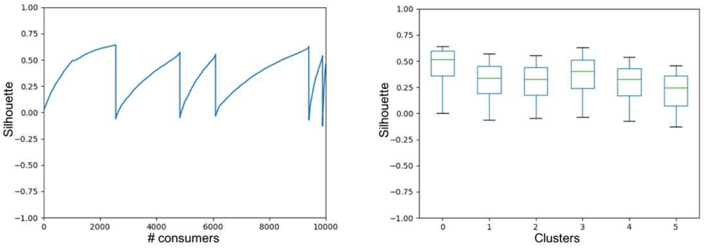
TABLE I. CONFIGURATIONS USED FOR THE K-MEANS CLUSTERING AND RESULTING SILHOUETTE INDICATORS *ASI* AND *GSI*.

Id	Dataset	Normalization	*K*	Distance	*ASI*	*GSI*
C1	cut points_9	contract	6	dtw_d2	0.346	0.304
C2	cut points_9	contract	7	dtw_d2	0.337	0.245
C3	cut points_9	contract	6	dtw_d3	0.335	0.268
C4	cut points_9	contract	7	dtw_d3	0.325	0.255
C5	cut points_9	contract	6	Euclidean	**0.367**	**0.327**
C6	cut points_9	contract	7	Euclidean	0.343	0.307
C7	cut points_9	max-min	6	Euclidean	**0.294**	**0.385**
C8	cut points_9	max-min	7	Euclidean	0.273	0.352
H1	hourly_240	contract	6	Euclidean	-0.015	-0.016
H2	hourly_240	contract	7	Euclidean	**0.421**	**0.060**
H3	hourly_240	contract	6	dtw_d10	**-0.082**	**-0.089**
H4	hourly_240	contract	7	dtw_d10	-0.060	-0.063

C. Details on the clustering results

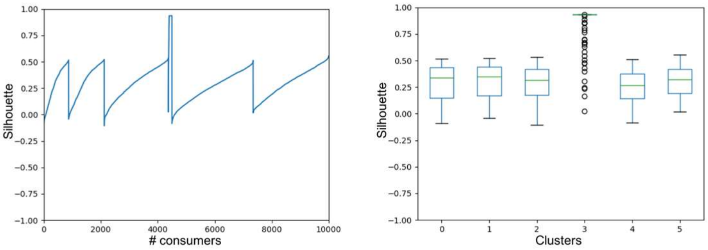
Let us look in more details at the results of a subset of cluster sets associated to configurations C5, C7, H2 and H3. Table II reports for each cluster the corresponding number of consumers, for each experiment. The number of clusters is six, except for Configuration H2, where is seven. Configuration H3 includes a homogeneous partition in terms of cluster cardinality. Configuration C5 includes 3 large-sized clusters and 3 medium-sized clusters, while configuration H2 presents a high data fragmentation among different clusters (i.e., 4 over 7 clusters include only a few number of consumers).

Fig. 6 shows the hourly consumption boxplot distribution for each cluster. In more detail, Fig. 6a shows the boxplot distribution of the hourly consumption with respect to the 9 input features, for configuration C5 separated for each cluster. The clusters are well partitioned highlighting different trends for each cluster. Fig. 6b and Fig. 6c show the first 24 hours of the 240 totals for configurations H2 and H3, respectively. Clusters that include only a limited number of consumers are those that represent a small number of outliers, as shown in Fig. 6b. While it is apparent that the clustering model shown in Fig. 6c is not good due to the presence of several outliers for each input feature, so that the result is not appropriate. The C5 partitioning is much more effective than the ones identified in configuration H2 and H3, highlighting a reduced number of outliers.



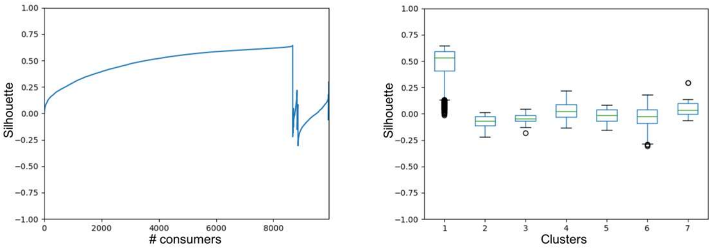
a) Silhouette plot for each consumer. b) Silhouette box plot for each cluster.

Fig. 2. Silhouette plots for configuration C5.



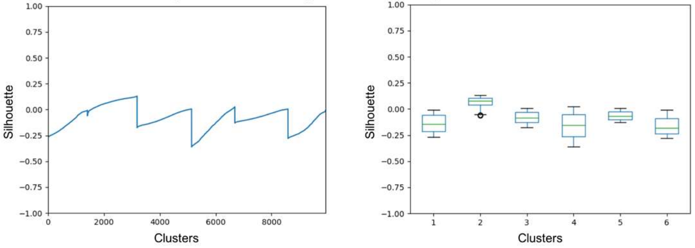
a) Silhouette plot for each consumer. b) Silhouette box plot for each cluster.

Fig. 3. Silhouette plots for configuration C7.



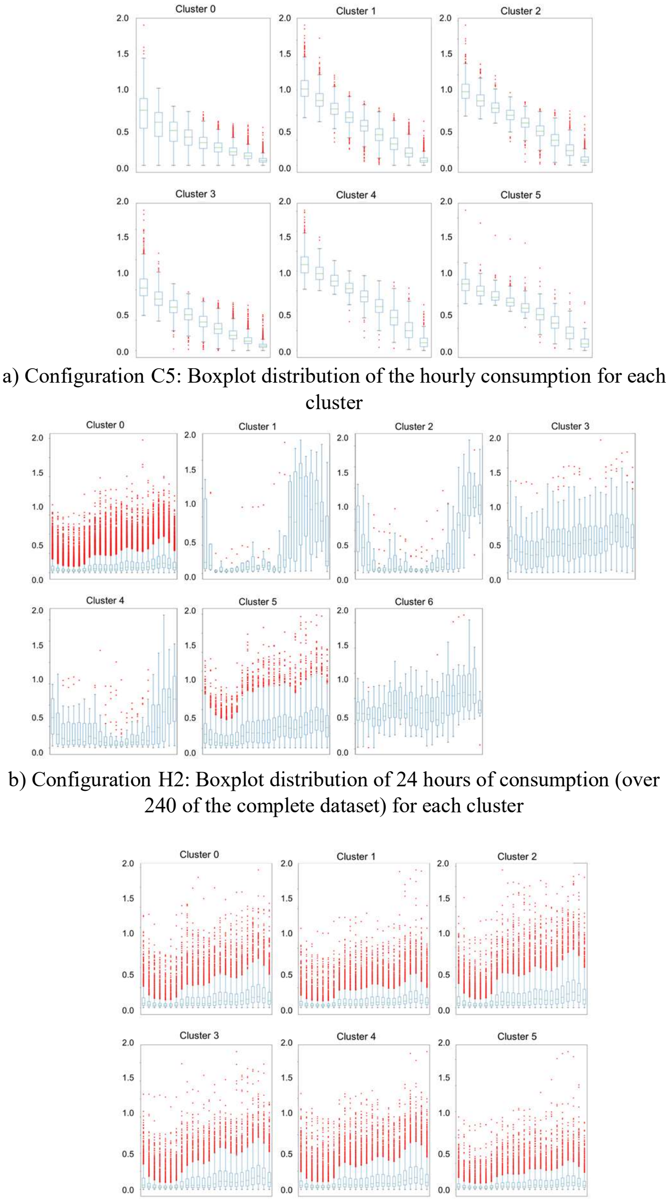
a) Silhouette plot for each consumer. b) Silhouette box plot for each cluster.

Fig. 4. Silhouette plots for configuration H2.



a) Silhouette plot for each consumer. b) Silhouette box plot for each cluster.

Fig. 5. Silhouette plots for configuration H3.



a) Configuration C5: Boxplot distribution of the hourly consumption for each cluster

b) Configuration H2: Boxplot distribution of 24 hours of consumption (over 240 of the complete dataset) for each cluster

c) Configuration H3: Boxplot distribution of 24 hours of consumption (over 240 of the complete dataset) for each cluster

Fig. 6. Boxplot distribution of the hourly consumption for each cluster.

TABLE II. NUMBER OF CONSUMERS IN EACH CLUSTER FOR DIFFERENT CONFIGURATIONS.

cluster	C5	C7	H2	H3
0	2554	869	8677	1409
1	2274	1250	8	1780
2	1265	2250	16	1950
3	3307	127	127	1548
4	486	2849	33	1904
5	114	2655	1063	1344
6	-	-	11	-

Fig. 7 shows the scatter plot that models the relations between the maximum and the average normalised hourly energy consumption, separately for each cluster. Different colours have been used to model cluster membership for each consumer. The C5 plot (top) indicates a very low overlap between the points belonging to the different clusters, while the overlap is much more evident for configuration H3 (bottom). Configuration H2 (middle) is a trade-off between the previous cases. However, the imbalance partitions are well depicted, showing the presence of clusters with fewer consumers. Although the experiment related to configuration C5 does not contain explicitly neither the average nor the maximum normalised hourly energy consumption, the identified partition provides separated clusters, much better than the partition identified through configuration H2 and configuration H3.

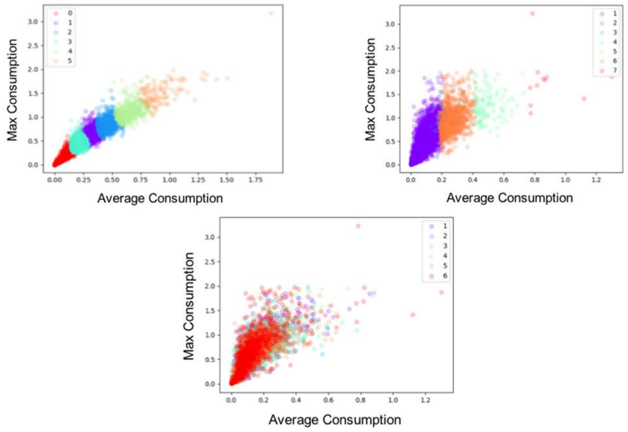


Fig. 7. Scatter plot with maximum and average normalised hourly energy consumption for configurations: C5 (top-left), H2 (top-right), and H3 (bottom).

V. Conclusions

This paper has presented the results of an extensive analysis carried out on the hourly data provided by the smart meters of real residential consumers in Spain. The key messages are:

- There is no generalised solution for the selection of the best features and parameters. The entire selection process has to be data-driven. The variants considered for the pattern selection, data normalisation, distance and other specific parameters have to be studied in a dedicated way for each type of dataset.
- For residential consumers, the choice of the method of analysis has to be carried out by taking into account the *curse of diversity* introduced by the presence of many peaks appearing at different timings, caused by the different consumers' behaviour (even of the same consumer in different days) in the usage of their appliances. For this purpose, the usage of 240 hourly data in a time window of two weeks has produced worse results than the usage of the 9 selected cut points taken from the normalized duration curve of the 240 hourly data.
- The data reduction technique that forms the selected set of cut points, based on the normalised duration curves constructed from the hourly data gathered in the time window of analysis, has provided very good results with respect to the state-of-the-art methods. This result is particularly interesting because the CONDUCT methodology allows identifying a very good quality partition, as demonstrated through the analysis of the Silhouette-based indices, by compactly representing the time windows with a few interesting points. This good result indicates that long residential consumption time series must be properly elaborated before mining them to effectively discover interesting insights from data.
- The analysis of the results has also highlighted that the numerical values of the clustering validity indicators have to be interpreted and not just taken as they stand to create a ranking. In fact, a simple *ASI* ranking would promote the configuration H2 as the best one, notwithstanding this configuration creates both a large consistent group and all the other poorly clustered groups. In this case, its low *GSI* indicator confirms the lack of effectiveness of configuration H2 with respect to the proposed configuration C5.

The research results discussed in this paper can be exploited in real-life settings to correctly characterize patterns of energy consumption (including the seasonality of consumption) and customer behaviour. This knowledge can be exploited to (i) effectively support a new data analytics task such as the energy consumption prediction; (ii) derive ad-hoc marketing strategies to satisfy real consumer behaviour and needs; and (iii) define guidelines on the consumers profiles more attractive to fill any gap between supply and demand.

References

[1] Di Corso, E., Cerquitelli, T., Apiletti, D. (2018). METATECH: METeorological Data Analysis for Thermal Energy CHaracterization by Means of Self-Learning Transparent Models. *Energies*, 11(6), 1336.

[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] Motlagh, O., Paevere, P., Hong, T.S., & Grozev, G. (2015). "Analysis of household electricity consumption behaviours: impact of domestic electric generation". *Applied Math. and Computation*, 270, 165-178.

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

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

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

[7] Cerquitelli, T., Chicco, G., Di Corso, E., Ventura, F., Montesano, G., Del Pizzo, A., Mateo González, A., & Martin Sobrino, E. (2018). "Discovering electricity consumption over time for residential consumers through cluster analysis". *Proc. 14th DAS*, Suceava, Romania, 24-26 May 2018.

[8] Chicco, G. (2012). "Overview and performance assessment of the clustering methods for electrical load pattern grouping". *Energy*, Vol. 42(1), pp. 68–80.

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

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

[11] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Xin, D. (2016). "Mllib: Machine learning in apache spark". *The Journal of Machine Learning Research*, 17(1), 1235-1241.

[12] Montero, P., & Vilar, J.A. (2014). "TSclust: An R package for time series clustering". *J. of Statistical Software*, 62(1), 1-43.

[13] R Core Team (2013). *R: A language and environment for statistical computing*.

[14] Giorgino, T. (2009). "Computing and visualizing dynamic time warping alignments in R: the dtw package". *J. of Statistical Software*, 31(7), 1-24.

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