Categorisation of Low-Voltage Three-Phase Electricity Users

(Article begins on next page)

Categorisation of Low-Voltage Three-Phase Electricity Users

Gianfranco Chicco
Dipartimento Energia "G. Ferraris"
Politecnico di Torino
Torino, Italy
gianfranco.chicco@polito.it

Daniele Bonansinga
Dipartimento Energia "G. Ferraris"
Politecnico di Torino
Torino, Italy
daniele.bonansinga@studenti.polito.it

Pietro Colella
Dipartimento Energia "G. Ferraris"
Politecnico di Torino
Torino, Italy
pietro.colella@polito.it

***Abstract*— The categorisation of the electricity users is carried out by using statistical analysis of the available data over a period of analysis, together with clustering techniques to create appropriate groups of consumers. The grouping is affected by the selection of the macro-categories based on general user's attributes. This paper introduces the novel notion of internal categories for each macro-category. The main aspects referring to the selection of the user internal categories are analysed within an integrated framework that uses two-dimensional features (i.e., reference power and monthly energy) and a clustering algorithm with a consensus-based procedure for the final categorisation. The partitioning into internal categories enables reducing the number of users to be considered together for load profiling purposes. The results are presented considering real data obtained over one year for a large set of low-voltage three-phase electricity users.**

Keywords—Electrical demand, load curve, clustering, statistical analysis, customer, internal categories.

I. INTRODUCTION

A. Motivation and Background

In the analysis of the electricity consumption, the categorisation of the users is a relevant aspect for the creation of customer groups that have a consistent electrical behaviour. This activity is carried out by gathering data from the field and continues until the load profiles become available for selected classes of users. The load profiles are load curves that indicate the representative behaviour of the electrical consumption for all the users that belong to the same class of users during a selected period of time and in specified loading conditions (e.g., weekdays or weekends, or with other seasonality effects) [1].

Load profiling is a peculiar activity for the retailers that operate in a competitive context. Besides the applications that refer to retailing practices, load profiling has various applications relevant for the distribution system operator (DSO) [2]. The DSO has access only to limited information concerning the users. In particular, the DSO has no access to the contract data that the users have with their retailers, nor to the private data of the users referring to their internal usage of electricity. The main accessible data are the voltage level and the reference power. The monthly energy (in some cases, only the annual energy) could be available as well. In a few cases, generally for large users, the DSO measures the power curves with a given time step (e.g., 15 minutes).

In the distribution systems, the load profiles can be used to compensate the lack of information on the users, when the data on the users are needed for executing power flow calculations for different purposes, such as state estimation [3]. Applications to mapping the distribution system feeders based on the monthly load profiles calculated from hourly load patterns are presented in [4] considering a feeder with initial unknown sectors and in [5] to determine the typical load profiles for each month. A load modelling-based approach, associated with the detection of voltage variation events in the distribution network to construct the features of the loads, is introduced in [6].

Fig. 1 summarises the main points of the procedure considered for creating the load profiles. In particular:

- *Definition of the macro-categories*: the macro-categories are formed by considering different types of users for retail purposes (e.g., residential, industrial, commercial, traction, lighting), or for DSO purposes (e.g., ordinary use, or temporary use), also considering the voltage supply level (mostly relevant for the DSO).
- *Definition of the internal categories*: for each macro-category, further internal categories can be identified based on available data such as the reference power and the energy metered for given periods (e.g., annual, monthly), depending on the data accessible to the retailers or the DSO. The reference power is the contract power, which also defines the nominal power of the meter.
- *Definition of the data sampling*: even in the presence of a new generation of smart meters with better resolution in time (in general, 15 minutes or 30 minutes) with respect to hourly data, the elaboration of all the data gathered continuously for all the users may be excessively demanding. For this purpose, statistical methods, such as the stratified sampling approach [7], are useful to establish the number of users to be monitored. The partitioning given for the internal categories is considered to form the layers. The stratified sampling approach is used for determining the minimum number of users to monitor for constructing the load profiles for each category of users considered [8], using the same statistical representativeness of the results (i.e., the same confidence level for each layer).
- *Choice of representative days & loading conditions*: once the minimum number of users to monitor has been defined, the data are gathered from the smart meters for a number of users equal to or higher than the minimum number. The metered data are then partitioned by considering the relevant periods (e.g., season), and the weekdays, weekends or anomalous days, to obtain a number of load patterns that can be associated to similar loading conditions, i.e., the load patterns which represent a number of days that can be pre-processed together to reduce the number of data to send to the following steps.
- *Bad data detection*: During the data pre-processing, the bad data have to be identified based on the absence of the data, the indication of unreliable data provided by the data gathering system, and the check of possible anomalies in the data, carried out by using specific expertise in the analysis of electrical load patterns (e.g., excessive values in the context of the user or meaningless negative values).

The bad data can be simply removed (if a sufficient number of data is already available for the same loading condition) or can be replaced with an artificial meaningful data determined through specific techniques [9].

- *Formation of the representative load patterns*: for each loading condition, using the available data and considering the possible absence of some data as indicated in the above point, the representative load pattern (RLP) is constructed through a weighted averaging process. The RLP can be represented either as it stands (e.g., with no normalisation) or can be normalised by using a suitable normalising factor (e.g., the reference or contract power, the average power, or the peak power). In the case of normalisation, the normalising factor has to be stored together with the normalised load pattern to allow for successive reconstruction of the load patterns that indicates the actual power.
- *Selection of the clustering method*: the RLP data can be directly used as input features of the clustering algorithm, or data reduction techniques can be applied to transform the RLPs into a different set of H features in the time domain, or in different domains [10]. The reduction in the dimension of the dataset has also been addressed for data visualisation purposes [11]. One or more clustering algorithm can be selected for grouping the users based on the selected features.
- *Execution of the clustering*: The clustering procedure is a mapping $\mathfrak{R}^{M,H} \rightarrow \mathcal{N}^{M,1}$ in which the input matrix $\mathbf{X} = \{x_{mh}\} \in \mathfrak{R}^{M,H}$ contains the $h = 1,\ldots, H$ features for each member $m = 1,\ldots, M$ of the input dataset, and the output vector $\mathbf{g} = \{g_m\} \in \mathcal{N}^{M,1}$ associates an integer value (i.e., the number of the cluster) to each member $m = 1,\ldots, M$ of the input dataset. In case of comparisons among different clustering techniques, the calculation of various clustering validity indicators may be used for providing a ranking among the algorithms. Independently of the features used in the execution of the clustering, the clustering validity indices have to be computed by using the same features (e.g., the RLPs) in all cases, with the different output vectors that characterise the output of each clustering execution.
- *Formation of the classes*: in general, the clustering run with different algorithms and features forms different groups. The groups of users obtained do not necessarily represent all the users with the same characteristics. A post-clustering check is carried out to identify possible refinements of the groups, based on the most recurrent attributes in each cluster. The clusters are then re-grouped, forming the class representative load patterns (CRLPs) by averaging the RLPs of the users that belong to the same class (a weighted average is needed if normalised RLPs are used, considering the RLP normalising factor as the weighting factor). The final number of classes could be different with respect to the number of clusters.
- *Formation of the load profiles*: the CRLPs are the final load profiles that represent each class in the specified loading condition.
- *Classification*: a classification phase is included at the end of the procedure, with the aim of identifying the user's class that may be attributed to a new user, or to a user that has changed its main characteristics [12].

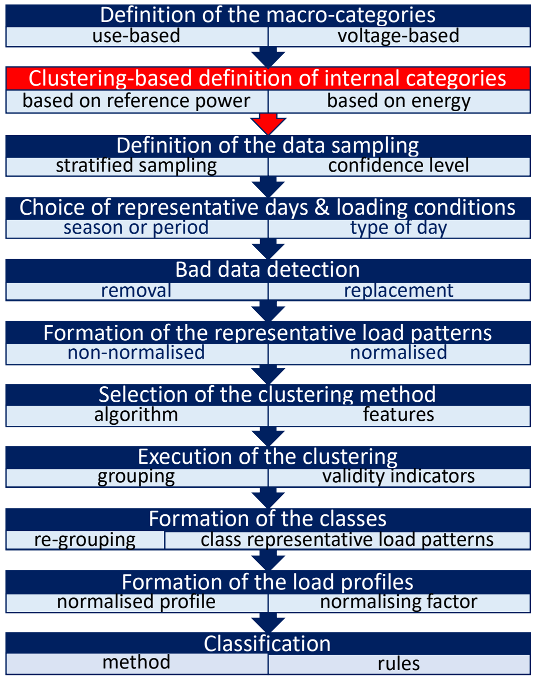

Fig. 1. Overall scheme of the activity that leads to load profiling, with highlight of the novel contribution of the paper.

More recently, the classical approach used for load profiling is being extended to the presence of active users (or prosumers), leading to revisiting the existing procedures in the light of a wider prosumer profiling [13].

Following the recent trends to apply more data-driven approaches based on the many data that become available from the new generations of smart meters, the classes formed and the corresponding load profiles could be updated more frequently than in the past. These updates may also be needed if the behaviour of one or more users changes based on the information available from a continuous flow of data.

B. Contributions and Organisation

The novelty of this paper is the introduction of the *internal categories* in the overall scheme that leads to load profiling. This provides an original view on the creation of user categories for a given macro-category of users, preliminary to further analyses aimed at creating for each category a number of classes for load profiling purposes. The key assumption used in this paper is that the data are managed by the distribution system operator (DSO).

The specific aspects of this paper refer to apply clustering procedures for the identification of the internal categories based on two-dimension (2D) data (reference power and energy) used simultaneously, rather than on power or energy analysed separately. Using only power data, the internal categories depend on the rated power of the users but not on the electrical energy. Conversely, using only energy data there is no link with the size of the plant. The resulting categories are different in the two cases. The use of 2D data enables a unique categorisation that considers both aspects.

The main points addressed are the preparation of the data to be used in a consensus clustering procedure specifically applied to form the internal categories, different from the clustering algorithms used in the next stage of formation of the classes based on the load patterns. Possible alternatives for creating the features used in the clustering procedures are discussed, in particular referring to data normalisation.

The data analysis is illustrated from the point of view of the expert of the electrical domain, rather than going into detailed comparisons among different clustering algorithms. Nevertheless, specific indications are provided on the usage of the algorithms for carrying out the calculations.

The next sections of the paper are organized as follows. Section II summarises the types of data used in the clustering procedure applied for the definition of the internal categories. Section III describes the consensus clustering procedure executed to aggregate into a single output the results of the individual clustering run for each month with the reference power and monthly energy data. Section IV shows the results of the application of the proposed procedure for defining the internal categories to a real large DSO dataset. The last section points out the concluding remarks.

II. Data Preparation for Clustering

The dataset considered refers to a subset of users supplied by the same DSO. A preliminary selection of a macro-category is carried out, on the basis of the internal rules adopted by the DSO to make a distinction among different types of applications, e.g., based on the voltage level, single-phase or three-phase supply, and continuous or temporary users. These distinctions do not contain information on the type of consumer, e.g., residential, industrial or commercial.

The lack of the information on the type of consumer is a significant difference with respect to what happens for the categorisation of the customer groups based on the information available from retailers.

The available data include the reference power and the energy for each month, making it possible to check the possible permanence of the same reference power at each month, as well as to consider the monthly energy consumption to carry out a separate analysis for each month.

If the reference power and the monthly or annual energy are used separately as clustering features, the groups formed will be different. In this paper, to avoid these different results, 2D features are used for clustering, so that visualisation of the results in two dimensions is possible. However, the presence of these two different features raises the issue of how to represent the features in the relative way, namely, the data may be used with their quantities (e.g., kW and MWh), or can be normalised with respect to their maximum value.

The choice on whether or not to apply normalisation affects the way the clustering operates and has a direct impact on the nature of the results. Specific examples are provided in Section IV to confirm this statement.

For a monthly analysis, the selected clustering algorithm is executed for each month. The choice of the clustering algorithm depends on various factors, among which scalability is a key factor when the number of users is very high. Then, concerning the features there are again the two options of using the non-normalised or the normalised features for the input data. Moreover, once the clustering results have been obtained for each month, a synthesis is needed to reach a final grouping of the users. For this purpose, a *consensus clustering* procedure has been developed in this paper to exploit the monthly information and provide a general result. The details are indicated in the next section.

III. Consensus Clustering

A. Procedure to implement the consensus clustering

The general scheme for consensus clustering is based on aggregating multiple clustering results, with three steps:

1) The procedure starts from the results of the selected clustering method, which is executed separately for each month. The reference power and monthly energy data for each month $j = 1, \ldots, J$, are contained in the vectors $\mathbf{g}_j$, which form the matrix $\mathbf{G} = \{\mathbf{g}_1 \ldots, \mathbf{g}_J\} = \{g_{mj}, j = 1,\ldots, J, m = 1,\ldots, M\}$.
2) Similarity indices are computed for each pair of users as shown in Eq. (1)

$$p_{a,b} = \sum_{j=1}^{J} k_f \qquad (1)$$

where:

- $p_{a,b}$ is the similarity index between the users denoted with the letters a and b;
- k_j is equal to $1/J$ if the users a and b belong to the same cluster for the month considered, and zero otherwise.

The similarity indices are collected in the symmetric similarity matrix $\mathbf{P} = \{p_{a,b}\} \in \mathfrak{R}^{M,M}$. The entry $p_{a,b}$ of the matrix $\mathbf{P}$ represents the posterior probability that the users a and b are located on the same cluster in all the months considered. The diagonal entries of the matrix $\mathbf{P}$ are equal to unity.

3) In the third step, with the matrix $\mathbf{P}$ as input, the users are merged into the desired number of groups running a spectral clustering algorithm, as described below.

B. Spectral clustering

The spectral clustering [14] considers the structure of the problem under analysis as a graph, with vertices and edges, and constructs the Laplacian matrix $\mathbf{L} \in \mathfrak{R}^{M,M} = \{\ell_{a,b}\}$, with the following characteristics:

- The off-diagonal terms are null if the users a and b never belong to the same cluster, otherwise, they contain the opposite of the similarity index $p_{a,b}$, for $i, j = 1, \ldots, M$:

$$\ell_{a,b} = -p_{a,b} \qquad (1)$$

- The diagonal terms contain the sum of the similarity indices located in the same row, e.g., for the user a:

$$\ell_{a,a} = \sum_{i=1}^{M} p_{i,a} \qquad (2)$$

Following the classical notation [15], the diagonal terms are arranged into a diagonal matrix $\mathbf{D} \in \mathfrak{R}^{M,M} = \mathrm{diag}(d_{i,i})$, and the Laplacian matrix is expressed as:

$$\mathbf{L} = \mathbf{D} - \mathbf{P} \qquad (3)$$

- Starting from the matrix $\mathbf{L}$, the spectral clustering algorithm computes the K eigenvectors that correspond to the K smallest eigenvalues and forms the matrix $\mathbf{U} \in \mathfrak{R}^{M,K} = \{u_{i,k}\}$ of which the K eigenvectors are the columns.
- The rows of the matrix $\mathbf{U}$ are formed as feature vectors (row vectors) $\mathbf{y}_i \in \mathfrak{R}^{1,K} = \{y_{i,k}\}$, for $i = 1, \ldots, N$.
- The feature vectors are sent to a clustering algorithm to form the clusters. The classical choice is the kmeans algorithm, with the initialisation of the centroids executed with the kmeans++ approach [16] to limit the effects of the randomness applied to the initial centroid selection – a typical drawback of the classical kmeans algorithm.

IV. Application to a Dataset of Low-Voltage Three-Phase Users

A. *Description of the dataset*

Let us consider the data extracted from the database of an Italian DSO for one year, referring to Low-Voltage three-phase users. The total number of users is 73,518, of which 46,396 belong to the most numerous macro-category called *ordinary use*. The analysis presented in this paper refers to the ordinary use macro-category only. The same reasoning can be applied to the other macro-categories. A possible exception are the low-voltage single-phase users, about 95% of which have basically the same contract power, then the contract power feature is poorly useful to partition the groups.

The first data filtering has reduced the data to 45,420 users with data present for all the 12 months of the year. Then, the number of users has been further reduced by removing the 127 users that changed the reference power during the year and further 11 users with inconsistent data (e.g., null reference power or excessive annual energy/power ratio). The remaining 45,282 users have been processed for determining their partitioning into a given number of groups based on their reference power and monthly energy data.

Several tests have been carried out, considering one-dimensional data with reference power only or energy only, and the 2D data with reference power and energy. The latter case is illustrated below. The kmeans clustering algorithm has been executed to form the groups. The rationale for the choice of the kmeans algorithm is its computation time, which has been found to be viable considering the size of the data processed (about 0.13 s for a single execution on a MacBook Pro laptop with 2.3 GHz Intel Core i9 8 core, with Matlab-coded solver). The initialisation of the centroids has been carried out by using the kmeans++ procedure [16]. For the size of the dataset used, other classical clustering algorithms such as the hierarchical clustering are affected by scalability problems and have not been tested, even though research for developing scalable versions of the hierarchical clustering algorithm is in progress [17].

Concerning the choice of the 2D features used for clustering, the following variants have been tested:

a) Clustering with reference power and annual energy data, executed with normalised and non-normalised features. The annual energy has been calculated by summing up the monthly energy during the year. The reference power is the same for each month by construction of the dataset.

b) Clustering with reference power and monthly energy data, executed for each month with normalised and non-normalised features. The results obtained from the application of the consensus clustering are also presented.

The exemplificative solutions shown in this section have been calculated by considering $K = 10$ clusters. This number of clusters has been chosen to represent a reasonable number of partitions that may be useful to the DSO for creating the main user categories, leaving to further analyses the creation of more detailed groups inside each user category.

B. *Clustering with reference power and annual energy data*

Fig. 2 shows the 2D mapping of the kmeans clustering results obtained with *non-normalised* features. Table I shows the corresponding number of users belonging to each cluster.

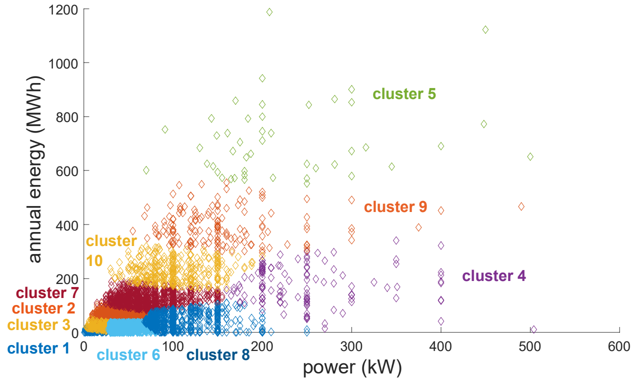

Fig. 2. Clustering results with $K = 10$ clusters with non-normalised reference power and annual energy input data.

Table I. Number of Users in the $K = 10$ Clusters (Non-Normalised Input Data for Reference Power and Annual Energy)

cluster	#users	cluster	#users
1	27889	6	3394
2	2647	7	877
3	9123	8	699
4	122	9	152
5	46	10	333

Fig. 3 shows the two-dimension mapping of the kmeans clustering results obtained with *normalised* features. Table II shows the corresponding number of users results belonging to each cluster. The normalisation with respect to the maximum values of each feature implies some changes in the structure of the clusters, however, these changes are not drastically relevant for the users with high reference power and high annual energy, whereas more evident changes occur for low reference power and low annual energy.

The numbering of the clusters in Fig. 2 and Fig. 3 has been obtained automatically, and no attempt has intentionally been made to renumber the clusters, because the groups created in each cluster are different in the two figures. Considering the number of users, some similarities can be observed. For example, the most numerous clusters (cluster 1 in Fig. 2 and cluster 4 in Fig. 3) are located at the bottom left of the figures, the less numerous clusters (cluster 5 in Fig. 2 and cluster 9 in Fig. 3) are the ones with high annual energy, and so forth with the other clusters.

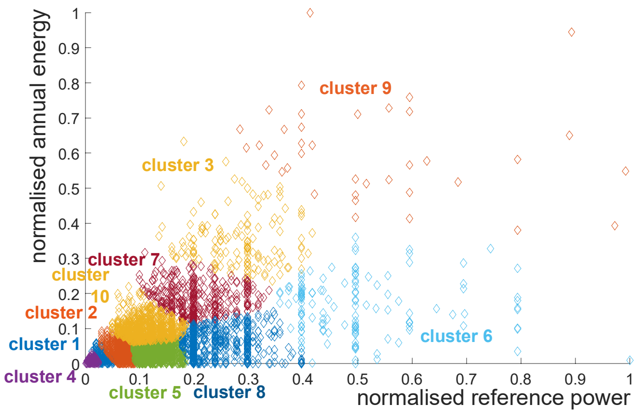

Fig. 3. Clustering results with $K = 10$ clusters with normalised reference power and annual energy input data.

TABLE II. NUMBER OF USERS IN THE $K = 10$ CLUSTERS (NORMALISED INPUT DATA FOR REFERENCE POWER AND ANNUAL ENERGY)

cluster	#users	cluster	#users
1	12400	6	138
2	5126	7	360
3	142	8	516
4	23774	9	38
5	1905	10	883

TABLE III. CONSENSUS CLUSTERING RESULTS FOR $K = 10$ CLUSTERS (NON-NORMALISED INPUT DATA FOR REFERENCE POWER AND MONTHLY ENERGY)

cluster	power range (kW)	#users	cluster	power range (kW)	#users
1	0-8	14711	6	40-41	15
2	9-11	9717	7	40-63	2054
3	12-21	11288	8	62-64	28
4	22-23	32	9	62-92	937
5	23.5-41	5440	10	> 89	1060

C. *Clustering with reference power and monthly energy data*

When monthly data are considered with the use of non-normalised reference power and monthly energy, the executions of the kmeans clustering algorithm provide the results shown in Fig. 4. The evident result is that the grouping is mostly based on the partition into successive ranges of reference power. The reason of this outcome is that the quantities represented in the two dimensions are different (the reference power is about one order of magnitude bigger than the monthly energy in MWh), so that the partitioning is mainly based on the identification of the reference power levels. Such a result is not surprising, as it normally happens when non-normalised patterns have different numerical values for each feature. For the same reason, if the monthly energy were represented in kWh rather than in MWh (with numerical values of two orders of magnitude bigger than the reference power values), the clustering results would show a "horizontal" partitioning of the users into monthly energy ranges (these results have been obtained during the testing and are not shown here because of their obvious meaning and for space reasons).

Starting from the individual monthly results, the consensus clustering has provided the results shown in Fig. 5. The numbers of the clusters are not indicated in the figure, and are sequentially reported in Table III, with very small overlapping of the reference power ranges. Moreover, the vertical scale of Fig. 5 reports the annual energy, for the sake of comparison with the points of Fig. 2.

An interesting aspect of the results obtained in this non-normalised case is that the partitioning of the reference power ranges is obtained by using information on both reference power and monthly energy, rather than on the reference power only. In this way, the definition of internal categories based on the reference power ranges is supported by a wide analysis based on combined features and consensus clustering.

When the normalised features are used for clustering, the results for $K = 10$ clusters are indicated in Fig. 6 for the 12 months considered individually (with values shown in kW and MWh for the reference pwer and monthly energy, respectively, even though the normalised values have been used in the clustering procedure), and in Fig. 7 for the results of the consensus clustering (with annual energy data indicated for consistency with Fig. 2 and Fig. 5). Moreover, Table IV reports the numbers of users for each cluster for the consensus clustering results.

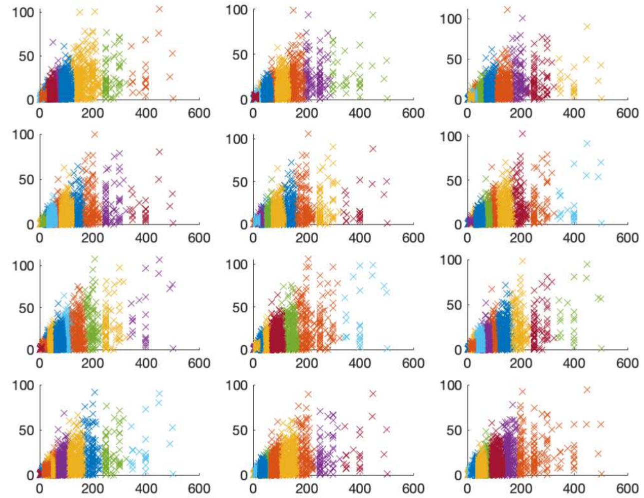

Fig. 4. Cluster partitioning for $K = 10$ clusters in the 12 months of the year with non-normalised reference power and monthly energy input data. The months are sequentially indicated by rows. Horizontal axis: reference power in kW; vertical axis: monthly energy in MWh.

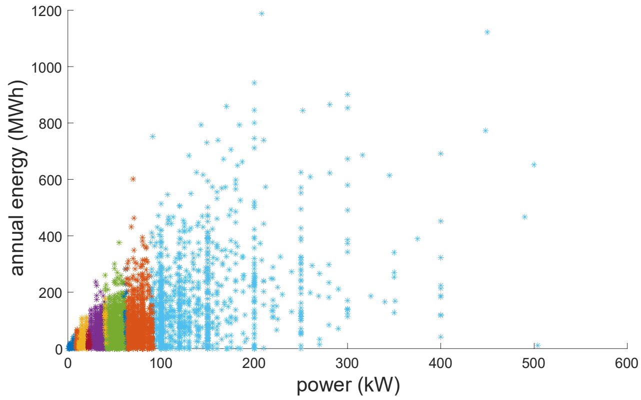

Fig. 5. Cluster partitioning from consensus clustering for $K = 10$ clusters with non-normalised reference power and monthly energy input data.

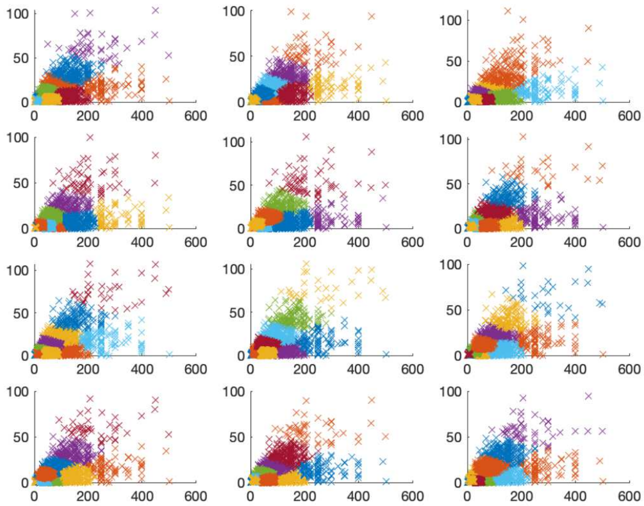

Fig. 6. Cluster partitioning for $K = 10$ clusters in the 12 months of the year with normalised reference power and monthly energy input data. The months are sequentially indicated by rows. Horizontal axis: reference power in kW; vertical axis: monthly energy in MWh.

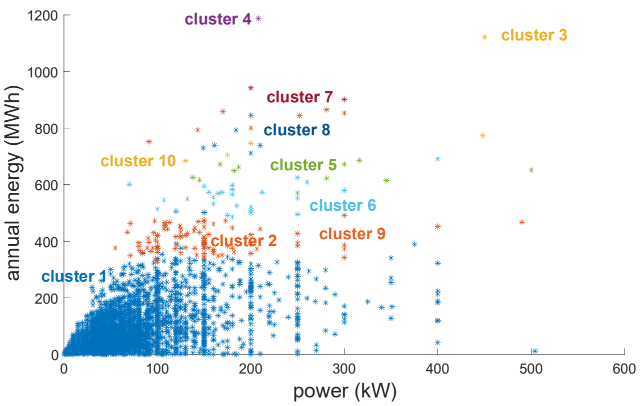

Fig. 7. Cluster partitioning from consensus clustering for $K = 10$ clusters with normalised reference power and monthly energy input data.

TABLE IV. CONSENSUS CLUSTERING RESULTS FOR $K = 10$ CLUSTERS (NORMALISED INPUT DATA FOR REFERENCE POWER AND MONTHLY ENERGY)

cluster	#users	cluster	#users
1	45130	6	28
2	7	7	2
3	1	8	6
4	1	9	92
5	11	10	4

The partitioning into clusters for each month is qualitatively similar to the partitioning shown in Fig. 3 with normalised reference power and annual energy data, as the normalisation keeps the values in the two dimensions at the same order of magnitude. However, the results of the consensus clustering indicate the creation of a very large group of users and the isolation of very small clusters as they were outliers. This effect depends on the permanence of many users in the same cluster of other users for various months, that creates a cross-link effect among many users with relatively low reference power and monthly energy, while the small clusters contain users with higher values of the features. In this last case, there is room for further improvement of the strategy used in the consensus clustering procedure, with the aim of avoiding the creation of a prevailing cluster that contains the majority of the users.

V. CONCLUSIONS

The creation of the user categories carried out by using reference power and energy data from the point of view of the DSO has some peculiarities that makes it different from the creation of user categories conducted with the point of view of the retailer. The main difference is the type of data and information available from the users. On these bases, in this paper a new stage that contains the creation of the internal categories for all macro-categories defined a priori has been explicitly included in the overall procedure for load analysis and profiling.

The main principles for conducting the creation of the internal categories have been presented, showing the results of applying some clustering procedures to a large dataset of low-voltage three-phase users. A specific contibution has been the formulation and application of a consensus clustering strategy for synthesising the results of individual clustering executed for each month into a single final result.

From the cases explored, data normalisation and the consensus clustering strategy have a relevant effect on the clustering results. When the reference power values are one order of magnitude higher than the energy values, the internal categories are formed with reference power ranges. This result could be useful to obtain a reference power partitioning based not only on reference power inputs.

The next step is to send the user partitioning based on reference power and annual energy (the most appropriate case obtained) to the stratified sampling of the internal categories, for determining the minimum number of users to monitor.

Further extensions include the development of a refined strategy for consensus clustering, the testing of different clustering algorithms, and parametric analyses on the number of clusters.

REFERENCES

[1] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer Characterization Options for Improving the Tariff Offer," *IEEE Trans. Power Syst.*, vol. 18 (1), pp. 381–387, 2003.

[2] D. Peters, R. Völker, F. Schuldt, and K. von Maydell, "Are standard load profiles suitable for modern electricity grid models?," *2020 17th International Conference on the European Energy Market (EEM)*, Stockholm, Sweden, 2020.

[3] G. Hong and Y.S. Kim, "Supervised Learning Approach for State Estimation of Unmeasured Points of Distribution Network," *IEEE Access*, vol. 8, pp. 113918–113931, 2020.

[4] G. Nourbakhsh, G. Eden, D. McVeigh, and A. Ghosh, "Chronological Categorization and Decomposition of Customer Loads," *IEEE Trans. Power Deliv.*, vol. 27 (4), pp. 2270–2277, 2012.

[5] M.A. Maniar and A.R. Abhyankar, "Two-Stage Load Profiling of HV Feeders of a Distribution System," *IEEE Systems Journal*, vol. 13 (3), pp. 3102–3110, Sept. 2019.

[6] X. Tang, K.N. Hasan, J.V. Milanović, K. Bailey, and S.J. Stott, "Estimation and Validation of Characteristic Load Profile Through Smart Grid Trials in a Medium Voltage Distribution Network," *IEEE Trans. Power Syst.*, vol. 33 (2), pp. 1848–1859, 2018.

[7] J. Neyman, "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection," *Journal Royal Statist. Society*, Part IV, pp. 558–606, 1934.

[8] G. Chicco, D. Labate, A. Notaristefano, and F. Piglione, "Unveil the Shape: Data Analytics for Extracting Knowledge from Smart Meters," *Energia Elettrica Supplement Journal*, vol. 96 (6), pp. 1–15, 2019.

[9] X. Chen, C. Kang, X. Tong, Q. Xia, and J. Yang, "Improving the Accuracy of Bus Load Forecasting by a Two-Stage Bad Data Identification Method," *IEEE Trans. Power Syst.*, vol. 29 (4), pp. 1634–1641, 2014.

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

[11] A. Aleshinloye, M.A. Manzoor, and A. Bais, "Evaluation of Dimensionality Reduction Techniques for Load Profiling Application in Smart Grid Environment," *IEEE Canadian Journal of Electrical and Computer Engineering*, vol. 44 (1), pp. 41–49, 2021.

[12] V. Figueiredo, F. Rodrigues, Z. Vale, and J.B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Trans. Power Syst.*, vol. 20 (2), pp. 596–602, 2005.

[13] G. Chicco and A. Mazza, "Load Profiling Revisited: Prosumer Profiling for Local Energy Markets," Chapter 13 in T. Pinto, Z. Vale and S. Widergren (Eds), *Local Electricity Markets*, Academic Press, pp. 215–242, 2021.

[14] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 22, pp. 888–905, 2000.

[15] U. Von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing Journal*, vol. 17 (4), pp. 395–416, 2007.

[16] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," *Proc. 18th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007*, New Orleans, LO, January 7-9, 2007, http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf

[17] N. Monath, A. Dubey, G. Guruganesh, M. Zaheer, A. Ahmed, A. McCallum, G. Mergen, M. Najork, M. Terzihan, B. Tjanaka, Y. Wang, and Y. Wu, "Scalable Hierarchical Agglomerative Clustering," arXiv:2010.11821v3 [cs.LG] 30 Sept. 2021.