Creation of the Internal Categories for Low-Voltage Three-Phase Electricity Users

Creation of the Internal Categories for Low-Voltage Three-Phase Electricity Users

Gianfranco Chicco, *Fellow, IEEE*, Daniele Bonansinga, Pietro Colella, *Member, IEEE*, and Lorenzo Solida, *Graduate Student Member, IEEE*

***Abstract*—The categorisation of electricity users is carried out by Distributor System Operators (DSOs) and retailers to create synthetic load profiles. Initially, some macro-categories are defined based on general attributes of the users, such as voltage level and type of users (residential/non-residential). Then, statistical analyses and clustering techniques are applied to create appropriate groups of users and compute the load profiles for each group. This paper considers the DSO point of view and proposes the definition of additional internal categories, determined by using mono-, two- and three-dimensional features based on reference power, energy, and utilisation. The internal categories contain subsets of users with similar characteristics, which are then sent to the subsequent steps of the procedure for computing the load profiles. The internal categories are formed by executing a clustering algorithm on annual data, or adopting a consensus-based procedure that evaluates the results of clustering algorithms executed for each month of the year. The results are presented considering real data obtained during the years 2019 and 2020 for a large set of low-voltage three-phase users. The effects of the severe restrictions due to the COVID-19 pandemic on some monthly energy data in 2020 are also considered in the analysis.**

***Index Terms*—Electrical demand, load curve, clustering, energy consumption, statistical analysis, customer category, COVID-19.**

I. Introduction

ELECTRICITY users are numerous and diverse. Their consumption changes by considering the type of user (e.g., residential vs. non-residential) and the specific way in which users manage their electrical devices. For residential users, the main aspects that determine their consumption are the composition and lifestyle of the individuals that act within the user's premises, leading to highly variable and poorly predictable individual power patterns. For non-residential users, the shapes of the power patterns of the users that belong to the same commercial category are generally different, because of the variety of equipment and management practices in place at each user. Therefore, the grouping of users cannot be based only on commercial categories.

Considering the above aspects, electricity users can be better categorised based on the shape of their power patterns, to provide meaningful information on how electricity is used. The goal is to create customer groups that have a consistent electrical behaviour. The categorisation activity starts with gathering data from the field. Subsequently, the classes of users are identified for a given number of classes and a load profile is determined for each class of users. The load profile is a power pattern that indicates the representative shape of electrical consumption for all the users that belong to that class during a selected time period and in specified loading conditions (e.g., weekdays or weekends, or with other seasonality effects) [1].

The load profiles can be used at the distribution system level, to compensate for the lack of information on the users when the users' data are needed for state estimation purposes [2]. Furthermore, typical load profiles can be determined for each month [3] and monthly load profiles calculated from hourly load patterns can be exploited for mapping the distribution system feeders considering that the initial partitioning of the users connected to the feeder is unknown [4]. The detection of voltage variation events in the distribution network can be associated with a load modelling-based approach to construct the features that represent the loads [5]. Recent developments refer to the creation of user groups in energy communities, with the aim of minimizing the reverse power flow at the point of supply of the distribution network, as well as shifting the peak demand of the system [6].

The categorisation of the electricity users can be seen from different points of view:

- For the retailer, load profiling is a peculiar activity for operating in a competitive context [7]. The growing deployment of smart meters with enhanced communication capabilities with retailers allows users to be more informed about their electricity usage and apply energy management, while retailers can set up customized retail price schemes for individual users [8]. Pricing signals can also be sent to selected users, identified through their load profiles, to participate in demand response programmes [9]. A demand response aggregator can interact with the wholesale electricity market by considering the attributes of the load profiles of aggregated loads [10].
- For the distribution system operator (DSO), the situation is different. The DSO has access only to limited information regarding the users. In particular, the DSO has no access to some contract data (e.g., electricity rates), nor to users' private data referring to their internal usage of electricity. The main accessible data are the voltage level and the reference power. Monthly energy may also be available (while sometimes there is only the information on the annual energy) [11]. In some cases, generally for large users, the DSO measures the power curves individually with a given time step (e.g., 15 minutes).

From both points of view, electricity users are categorised by carrying out statistical analysis of the data available over an observation period. The data is then managed to provide the

The authors are with Politecnico di Torino, Dipartimento Energia "Galileo Ferraris", Corso Duca degli Abruzzi 24, 10129 Torino, Italy (email: gianfranco.chicco@polito.it, dbonansinga97@gmail.com, pietro.colella@polito.it, lorenzo.solida@polito.it).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

inputs for executing suitable clustering algorithms aimed at creating appropriate user groups. To avoid the analysis of the entire data set available about all users, typically the users are partitioned into *macro-categories* based on general attributes. These general attributes might be different considering retailers or DSOs. For a retailer that operates on a specific type of users, for example, low-voltage users only, the main distinction could be based on residential vs. non-residential users and the type of contract applied. For the DSO, the main distinction could refer to users with Low/Medium/High Voltage supply, then also residential vs. non-residential users, and the type of grid connection (e.g., permanent or temporary installations).

In their previous work [12], the authors considered the DSO point of view to introduce the novel notion of *internal categories* for each macro-category. The definition of internal categories is based on the limited data available to the DSO and is functional for understanding the relative impact of each internal category on the overall consumption, over different periods of time. The rationale for the introduction of the internal categories is that the range of reference power values for three-phase low-voltage customers is quite wide (from a few kW to hundreds of kW) and the energy consumption of these customers is also highly variable. As such, it is hard to compare the electrical behaviour of these customers on a common ground in a relatively limited range of the relevant variables. A further distinction of the customers into internal categories is then helpful.

This paper is the extended version of the conference paper [12] in which further novel contributions have been added, with the following specific changes:

- The references to literature works are extended by including further contributions on different aspects, from the general context of electricity customer partitioning to the clustering techniques used in the specific context of analysing data from three-phase low-voltage electricity consumers.
- The discussion on the adoption of the reference power and the monthly energy consumption to form two-dimensional (2D) features is extended, with the introduction of the new *utilisation* feature to form new multi-dimensional features to be used as inputs for clustering, namely, 2D features together with the reference power or the monthly energy consumption, up to three-dimensional (3D) features formed by adding the utilisation to the other two entries.
- The discussion on the use of multi-dimensional features in the clustering process for creating the internal categories has been extended, based on the fact that the 2D and 3D features considered are not independent of each other. Results of comparisons carried out by using individual, 2D and 3D features are presented and compared.
- Two years of data are analysed, referring to the years 2019 and 2020. In particular, in the year 2020 there was a remarkable impact of the COVID-19 pandemic on the electricity usage [13]. During the lockdown period, an increase in consumption for residential users and a reduction in consumption for non-residential users were experienced [14][15]. The three-phase users considered are typically non-residential, with an expected energy reduction. The results are discussed to assess the impact of COVID-19 in the definition of the features used to form the internal categories, with respect to the pre-COVID-19 situation. For the year 2019, the internal categories are formed by taking the users present throughout the entire year. For the year 2020, specific insights come from analysing the results found during the lockdown period.
- The clustering algorithms have been executed by applying the kmedoids algorithm (instead of the kmeans algorithm in [12]), in such a way that the representative set of features for each cluster is taken from existing data rather than being a centroid formed by averaged data.

This paper is focused on data analysis addressed from the perspective of the electrical domain expert, rather than going into detailed comparisons among different clustering algorithms. The traditional kmedoids clustering algorithm is used, with effective results to handle the large data set considered.

More information on aspects relevant to the context of this paper may be found in [16], where the clustering techniques applied to smart grids are reviewed with a look to future trends, and in [17], which reviews distribution network-oriented applications that benefit from the analysis of the data collected from the smart meters.

The user data refers to the user load, without the presence of local generators. Any local generator should be analysed as a separate macro-category. Extensions of the classical approach used for load profiling to the presence of active users (or prosumers) would lead to revisiting the existing procedures in the light of wider prosumer profiling [18], which is beyond the scope of this paper.

The results are presented considering the real data obtained for a large set of low-voltage three-phase electricity users. The three-phase users are generally non-residential, with easier categorisation of the behaviour of individual users.

The next sections of this paper are organised as follows. Section II presents a schematic approach to load profiling, which includes the upgrade related to the introduction of internal categories. Section III discusses the selection of the features for clustering and the clustering algorithms used. Section IV shows an application to a large dataset of an Italian DSO. The last section contains the concluding remarks.

II. Upgraded General Load Profiling Scheme

Fig. 1 summarises the main points of the procedure considered for the creation of the load profiles, where the novel definition of the internal categories is highlighted with red background.

In particular, the individual points are:

- *Definition of the macro-categories*: the macro-categories are formed by considering different types of users (e.g., residential, industrial, commercial, traction, lighting), also considering the voltage supply level (especially relevant for the DSO).
- *Definition of internal categories*: for each macro-category, further internal categories can be identified based on available data such as reference power and energy metered for given periods (e.g., annual, monthly), depending on the data accessible to retailers or the DSO.

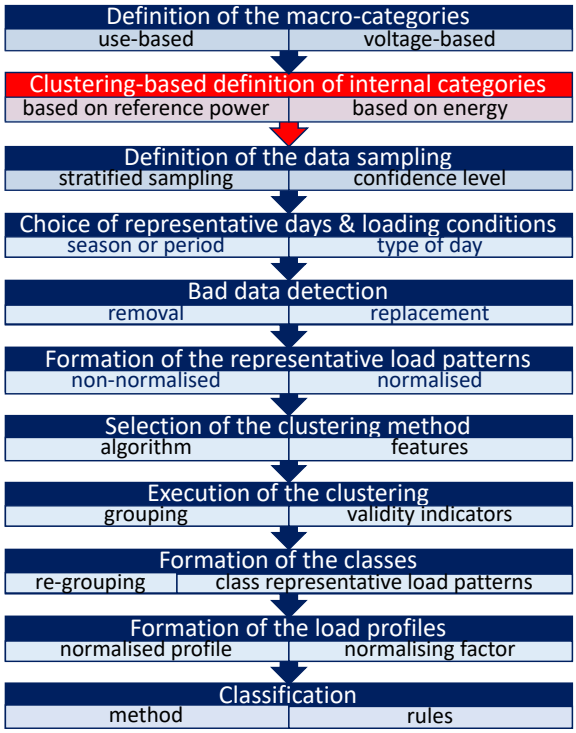


Fig. 1. Overall scheme of the activity that leads to load profiling.

- *Definition of the data sampling*: even in the presence of a new generation of smart meters with better time resolution (in general, 15 minutes or 30 minutes), the processing of all the data collected continuously for all the users may be excessively demanding. To reduce the computational time with the same statistical representativeness of the results, statistical methods are useful, such as the stratified sampling approach [19]. The partitioning given for the internal categories is considered to form the strata. In particular, the stratified sampling approach allows establishing the minimum number of users to monitor for constructing the load profiles for each category of users considered, with the same statistical representativeness of the results (i.e., the same confidence level for each stratum, based on the partitioning given for the internal categories) [20].
- *Choice of representative days and loading conditions*: once the minimum number of users to be monitored has been defined, the data is gathered by the smart meters for a number of users, randomly selected within the internal category, equal to or higher than the minimum number. The metered data are then partitioned by considering the relevant periods (e.g., seasons), and the weekdays, weekends or anomalous days, to obtain a number of load patterns that can be associated with similar loading conditions, i.e., the load patterns that represent a number of days which can be pre-processed together to reduce the number of data to be sent to the following steps.
- *Bad data detection*: During data pre-processing, data-polishing activities have to be carried out. Bad data can be either missing data or outliers identified using specific expertise in analysing 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 sufficient data is already available for the same loading condition) or can be replaced with meaningful artificial data determined through specific techniques [21].
- *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 previous point, the representative load pattern (RLP) is constructed through a weighted averaging process. The RLP can be represented either as it stands (e.g., without normalisation) or can be normalised by using an appropriate normalising factor (e.g., reference or contract power, average power, or peak power) [22]. In the case of normalisation, the normalising factor must be stored together with the normalised load pattern to allow subsequent reconstruction of the load patterns with 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 [23]. Reducing the dataset size has also been addressed for data visualisation purposes [24]. 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 the case of comparisons among different clustering techniques, the calculation of various clustering validity indicators can be used for providing a ranking among the solutions obtained from different algorithms. Regardless of the features used in the execution of the clustering, the clustering validity indices must be computed using the same features (e.g., the RLPs) in all cases, with different output vectors that characterise the output of each clustering execution.
- *Formation of the classes*: in general, the clustering executed with different algorithms and features forms different groups. The groups obtained do not necessarily represent all the users with the same characteristics (some users could be assigned to other clusters, as the clustering is not perfect). A post-clustering check is carried out to identify possible refinements of the groups, based on the most recurring attributes in each cluster. The clusters are then re-grouped, forming the final classes.
- *Formation of the load profiles*: the class representative load patterns (CRLPs) are formed by averaging the RLPs of the users that belong to the same final class (a weighted average is needed if normalised RLPs are used, considering the RLP normalising factor as the weighting factor). The CRLPs are the final load profiles that represent each final class in the specified loading condition.
- *Classification*: at the end of the procedure, a classification phase is included with the aim of identifying the user class that can be attributed to a new user, or to a user that has changed its main characteristics [25].

III. Feature Selection and Clustering Algorithms for the Definition of the Internal Categories

A. Feature Selection

From the DSO point of view, the macro-categories are based on the voltage supply level of the user (High/Medium/Low Voltage), which determines separate sections of the network. For the definition of the internal categories, the following individual features are of interest:

1. The *reference power* P_{ref}, that is, the contract power of the user, which corresponds to the setting of the general protections of the user's plant at the supply point. The reference power characterises the user and generally does not change, unless the user changes activity or needs a power upgrade; these cases must be monitored accurately, to avoid erroneous calculations when the reference power is considered together with other quantities.
2. The *electrical energy* $E_{\Delta T}$, which corresponds to the consumption in a given time period ΔT.
3. The *utilisation* $U = E_{\Delta T}/P_{\text{ref}}$, calculated in the time period ΔT (year or month) as the ratio between the actual electrical energy used and the reference power. The utilisation is expressed in hours and represents the equivalent number of hours at which a constant power P_{ref} would provide the same energy as the actual electrical energy consumption.

By definition, the above features are not independent of each other. The correlations between pairs of features can be determined by computing the correlation coefficients $\rho_{P,E}$, $\rho_{P,U}$ and $\rho_{E,U}$.

On these bases, a clustering algorithm can be executed in different ways and with different scenarios. For example, two main scenarios consist of:

1. Using the *annual* energy and utilisation, together with the reference power.
2. Using the *monthly* energy and utilisation, together with the reference power, separately for each month.

In both cases, it is then possible to:

a) Consider a *single feature* at a time, leading to different results in each case.
b) Consider *more features* at the same time, leading to the execution of a clustering with interdependent features, because of the definition of the features used, with a single result.

As discussed in [12], if the order of magnitude of the data is different, this may impact on the clustering outcomes. For easier comparison, the data should be normalised before running the clustering procedure. For the dataset under analysis, the power data are available in kW and the energy data in kWh. When the features used are in a single dimension (1D), there is no need for normalisation. Conversely, normalisation is more appropriate with features defined in 2D or 3D. In the latter cases, appropriate normalising factors have to be defined. Some intuitive choices for normalisation could be the maximum reference power $P_{\text{ref}}^{\max}$ for the power data, the number of hours in the month (e.g., 720 hours for a month with 30 days) for the monthly utilisation data, and the number of hours in the year (8760 hours for non-leap years, or 8784 hours for leap years) for the annual utilisation. Moreover, the energy data could be normalised by assuming the product between $P_{\text{ref}}^{\max}$ and the number of hours in the month (or in the year) as the normalising factor for the monthly energy (or the annual energy). However, for the dataset under analysis, these choices could not be the most appropriate ones. In fact, the distribution of the entries for the reference power, the energy and the utilisation factor have only a few entries close to the maximum values. For this reason, the normalising factors are chosen as the value that is not exceeded by 95% of the entries for each one of the features used (the practical aspects are shown in the application illustrated in Section IV.B).

The solution that uses annual data and the three normalised features together at the same time (in 3D) is taken as the benchmark case for this application, as it provides a single result. To prepare the data for this kind of analysis, it is essential to identify the presence of possible users that changed the reference power during the year and to exclude these users from the definition of the internal categories based on annual data. Moreover, these users are considered as separate entities before and after the change of reference power. The excluded users may be assigned a posteriori (i.e., after the definition of the internal categories) to the internal category that results most similar based on the selected features.

B. Clustering Algorithms

For clustering the data referring to the entire year or to individual months, the selected clustering algorithm has to be executed to form the groups. The classical kmeans algorithm used in [12], which minimises the within-cluster variances based on squared Euclidean distances [26], is a viable choice for clustering when the objective is to form relatively uniform groups rather than searching for uncommon individuals (outliers) and the computation time could be an issue. In the kmeans algorithm, the centroid of each cluster is computed by averaging the features of all the individuals in that cluster.

In this paper, with respect to [12] the clustering algorithm has been changed to kmedoids, in which the medoid of each cluster (i.e., the individual closest to the cluster centroid) is used to represent the cluster. In this way, the medoid is an existing individual (rather than the centroid, which is an averaged element different from the initial data). The objective function of kmedoids, to be minimised, is the sum of the distances from the individuals in each cluster to the medoid of the cluster. With large numbers of members in each cluster, the medoids are also relatively close to the centroids. A remarkable property of kmedoids with respect to kmeans is the robustness to noise and outliers, which is essential in the specific application, due to the large variability of the input data. The use of kmedoids is also motivated by the results of comparisons carried out among various clustering methods, as in the assessment of electrical energy consumption in a set of buildings [27].

On the other hand, the result of the kmedoids algorithm is not necessarily the global optimum [26]. As such, the clustering process is repeated for a given number of replicates with different initial medoids, and the result of the clustering is the one with the smallest objective function. Moreover, kmedoids can be applied with different methods. For M members of the dataset and K clusters, the classical partitioning around medoids (PAM) method has the important drawback of having a

computational complexity $O(K(M-K)^2)$, with respect to $O(M)$ for the kmeans algorithm. Therefore, the Matlab kmedoids solver [28] uses PAM only for M lower than 3000 and considers other options indicated as 'small' (for M variable from 3000 to 10000) and 'large' (for M higher than 10000), in which repeated searches are carried out with a kmeans-like update mechanism.

Considering the size of the data processed (tens of thousands of users, see Section IV.A), kmedoids (executed with the 'large' option available in Matlab, with 10 replicates) has been found to be viable (about 1 s for a single execution on a MacBook Pro laptop with 2.3 GHz Intel Core i9 8 cores). To limit the effects of the randomness applied to the initial centroid selection, the initialisation of the centroids (from which the initial medoids are identified) has been carried out by using the effective kmeans++ procedure [29]. For the size of the dataset used, other classical clustering algorithms could be affected by scalability issues (e.g., hierarchical clustering faces some issues in the construction of the pair-to-pair distance matrix among all users, even though research for developing scalable versions of the hierarchical clustering algorithm is in progress [30]).

C. Metrics for Assessing the Clustering Validity

The results of the clustering executed with different features are compared by using some metrics already applied to assess non-residential load curves [18]. These metrics includes the clustering validity indicators denoted as Mean Index Adequacy (*MIA*) [1], Clustering Dispersion Indicator (*CDI*) [1], Davies Bouldin Index (*DBI*) [31], Modified Dunn Index (*MDI*) (in the version with Euclidean distances adapted from [32]), and Ratio of within cluster sum of squares to between cluster variation (*WCBCR*) [33]. All indicators have the property that lower values correspond to better clustering validity [23]).

D. Determination of the Number of Clusters

The kmedoids algorithm requires the number of clusters as an input. In many clustering applications, the number of clusters is determined by executing specific procedures and indicators, identifying suitable conditions reached by the indicator when the number of clusters changes (e.g., with the classical elbow criterion). For this purpose, in this paper, the *CDI* and *WCBCR* clustering validity indicators are calculated in function of the number of clusters. In particular, the use of the *WCBCR* indicators has been suggested in the testing carried out in [34] because of the higher regularity of its variation in comparison with other indicators.

In the specific application presented in this paper, the clusters obtained are the classes to be used in the stratified sampling approach that follows the formation of the internal categories (Fig. 1). Therefore, the number of clusters should remain relatively limited (e.g., to about 10 clusters) to make the number of entries in the clusters statistically significant. Based on these aspects, a meaningful number of clusters for grouping the users is decided before executing the clustering algorithm.

E. Consensus Clustering Algorithm for the Aggregation of the Results Obtained for the Individual Months

The execution of the clustering procedures for the individual months creates different clusters at each month. These outcomes have to be merged for creating a single result based on the aggregation of the monthly results. For this purpose, a key issue appears in the dataset: the numbers of the clusters that are found by the clustering algorithm are different each time, and there is no relation between the numbers of the clusters and the location of the points. Because of that, it is essential to adopt a procedure that does not depend on the specific labelling used to assign the cluster identifiers and treats the cluster numbers as independent categorical variables.

The algorithm implemented, proposed in [12], is based on a customised application of the spectral clustering [35].

For each set of features considered, the customised procedure for consensus clustering includes three steps:

1) The procedure starts from the results of the clustering executed separately for each month. The features used 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:
$$p_{a,b} = \sum_{j=1}^{J} k_j \quad (1)$$
where:
   - $p_{a,b}$ is the similarity index between the users denoted with the letters a and b (i.e., the posterior probability that the users a and b are located on the same cluster in all the months considered);
   - 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 diagonal entries of the matrix **P** are equal to unity.
3) In the third step, using the matrix **P** as input, the users are merged into the desired number of groups running a spectral clustering algorithm, as described below.

The spectral clustering [36] 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 are never located in the same cluster, otherwise, they contain the opposite of the similarity index $p_{a,b}$, for $a = 1, \ldots, M$ and $b = 1, \ldots, M$:
$$\ell_{a,b} = -p_{a,b}$$
- The diagonal terms contain the sum of the similarity indices located in the same row, e.g., for the user a:
$$\ell_{a,a} = \sum_{i=1}^{M} p_{i,a} \quad (2)$$
  Following the classical notation [35], 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:
$$\mathbf{L} = \mathbf{D} - \mathbf{P} \quad (3)$$
- Starting from the matrix **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 **U** are formed as feature vectors (row vectors) $\mathbf{y}_i \in \mathfrak{R}^{1,K} = \{y_{i,k}\}$, for $i = 1, \ldots, M$.

To form the clusters, the feature vectors are sent to the kmeans algorithm, with centroids initialised with the kmeans++ approach.

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

A. Description of the Dataset

The data considered to provide the full details on the application presented in this paper are extracted from the database of an Italian DSO for the year 2020 and refer to low-voltage three-phase users. The total number of users slightly changes during the month, with an average of 74,090 and a variation from the maximum to the minimum of 205 users. Among these users, the most numerous macro-category is called *ordinary use* and is the one taken into account in this paper. The ordinary use category has an average number of 45,475 users and a slight variation (157 users) from the minimum to the maximum in the various months of the year.

Some filtering has been applied to exclude users without fully available and consistent data. 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 33 users with inconsistent data (e.g., null reference power, excessive annual energy/power ratio, or excessive monthly utilisation) or reference power higher than $P_{\mathrm{ref}}^{\max} = 500$ kW. The remaining 45,260 users have been processed for determining their partitioning into a given number of groups based on the three features (reference power, monthly energy, and utilisation). The overall number of data filtered is relatively low, and the dataset used for the calculations has a considerable dimension.

The calculation of the correlation coefficients between the pairs of features based on the annual data for each user results in $\rho_{P,E} = 0.721$, $\rho_{P,U} = 0.172$, and $\rho_{E,U} = 0.530$. Even though the three features are not independent with each other, there is a significant difference in the correlation coefficients. For example, the correlation between power and annual energy is remarkably higher than the correlation between power and annual utilisation. Because of that, using power and annual utilisation leads to less dependent features, which could be an interesting point for providing data with lower dependence to the clustering procedure.

The calculation of the correlation coefficients by considering the three features has been repeated with the monthly data. The results are shown in Fig. 2. The order of magnitude of the correlation coefficients remains similar during the months, with some reductions starting from March 2020, caused by the emergence of the COVID-19 pandemic (some insights on the related aspects are shown in the Appendix).

B. Normalising Factors

As anticipated in Section III.A, the normalising factors for the features are chosen as the 95% non-exceeding probability values for each feature, because only a few values of the features are located in the top 5% of the empirical Cumulative Distribution Function (CDF), and the variation that occurs in the top 5% values is very high. Fig. 3 shows the empirical CDFs for all the features. The normalising factors are then calculated as follows:

- for the reference power: 60 kW
- for the monthly energy: 5796 kWh
- for the annual energy: 67915 kWh
- for the monthly utilisation: 199 h
- for the annual utilisation: 2200 h

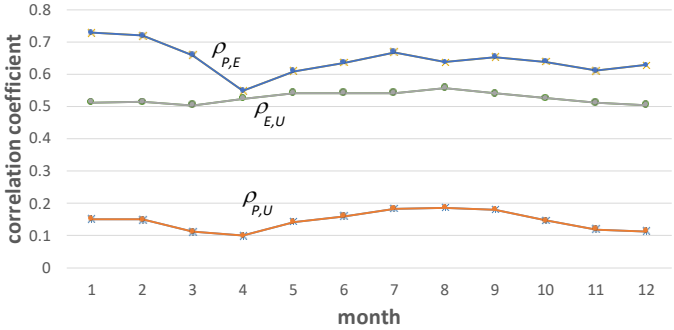


Fig. 2. Monthly correlation coefficients in the year 2020.

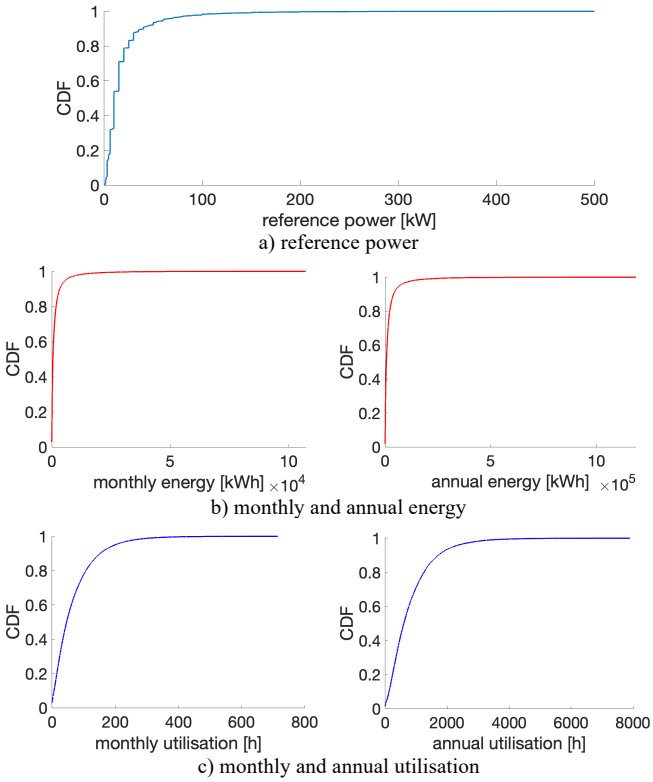


Fig. 3. Empirical CDFs of the features considered in the year 2020.

C. Clustering Results

For the sake of uniformity in the representation of the results obtained with the different variants of the 1D, 2D and 3D features used for clustering, the comparisons are shown by using a 2D plot with power (in kW) on the horizontal axis and energy (in MWh) on the vertical axis, also when the features used for clustering are different.

C.1. Definition of the Number of Clusters

For the definition of the number of clusters, the kmedoids algorithm has been executed with the reference power, annual energy and annual utilisation as individual features, for a number of clusters from 5 to 100, by tracking the *CDI* and *WCBCR* clustering validity indicators. With the elbow criterion, the results shown in Fig. 4 consistently indicate a number of clusters from about 10 to 15. Considering the aspects discussed in Section III.D about the need to keep the number of clusters relatively limited for the successive application to stratified

sampling, the number of clusters for the execution of the kmedoids clustering has been set to $K = 10$ clusters.

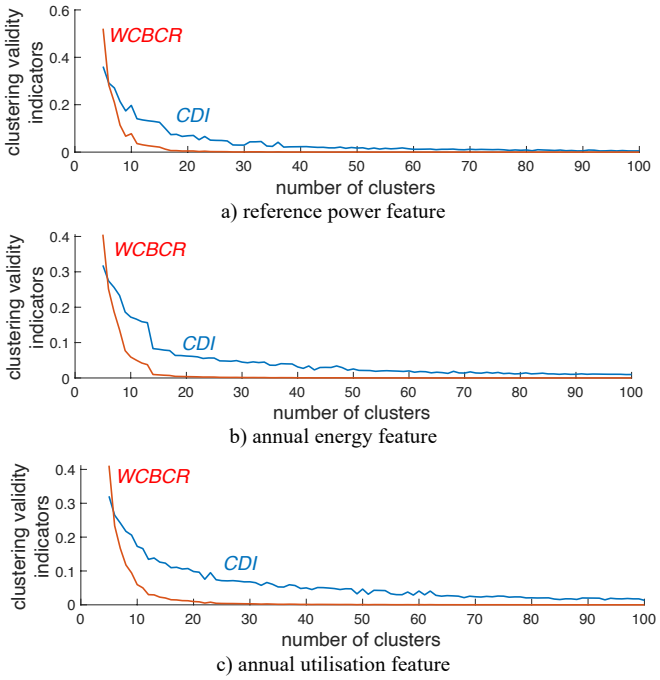


Fig. 4. Clustering validity indicators *CDI* and *WCBCR* for variable numbers of clusters considering different individual features (year 2020).

C.2. Analysis with Annual Data

The first example considers 1D clustering and the three features (power, annual energy, and annual utilisation) taken one at a time, for $K = 10$ clusters (Fig. 5). As expected, the internal categories are clearly defined based on the relevant feature. However, the results are very different with each other and do not account for the mixed aspects of the electricity usage. In particular, using only reference power data, the internal categories do not depend on the electricity consumption. Conversely, using only energy data there is no direct link with the size of the plant. Considering the utilisation as a feature takes into account both power and energy, however, losing the connection with the size of the plant.

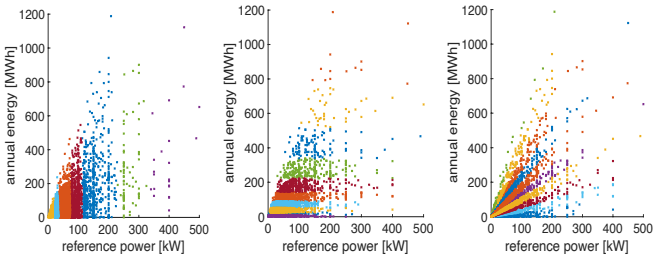


Fig. 5. Results of 1D clustering with annual data and different features.

The above aspects raise some interest in adopting multi-dimensional features. Fig. 6 shows the clustering results for $K = 10$ in the 3D case in which the reference power, annual energy and annual utilisation are assumed as features. The internal classes are formed with a higher variety in the cluster composition. The details on the location of the users in the clusters indicated in Fig. 7 further clarify the situation. The users with high reference power are grouped together in two clusters (the seventh cluster for the users with relatively high annual energy, and the ninth cluster for the users with relatively low annual energy). The effect of the annual utilisation is more evident for the users with very low power (the users in the fifth, second and first cluster have relatively higher annual utilisation, respectively detectable from the increasing slope of the trend of the group of points located in these clusters). Other groups are well-partitioned by the clustering algorithm.

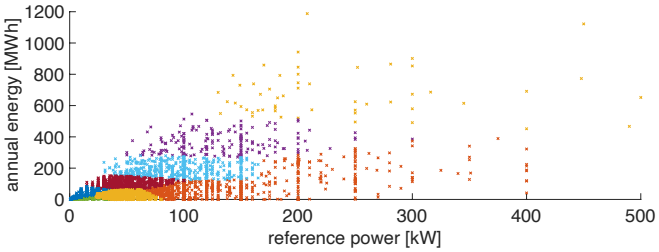


Fig. 6. Results of 3D clustering with annual data and the three features (reference power, annual energy, and annual utilisation).

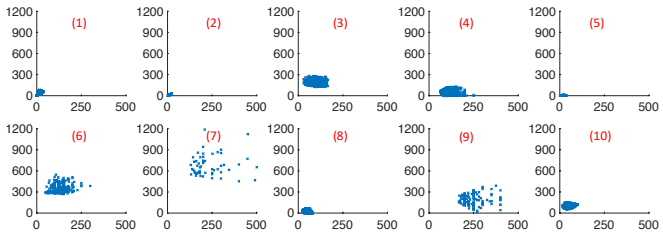


Fig. 7. The $K = 10$ clusters formed from the 3D clustering with annual data. Horizontal axis: reference power [kW]; vertical axis: annual energy [MWh].

To gain further insights in this direction, Fig. 8 shows the results of the 2D clustering executed on the annual data with the three combinations of pairs of features. The mixed effects of the pair of features used are evident. When the reference power and the energy are used as features, the clusters are formed by partitioning the users based on ranges of reference power and annual energy. When the annual utilisation is considered as a feature together with the annual energy or the annual utilisation, there is a slope-based contribution of the utilisation mainly in the clusters with lower power or lower energy, respectively.

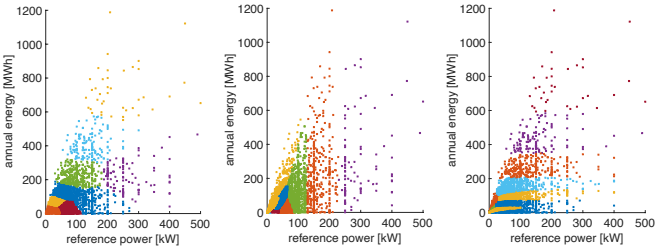


Fig. 8. Results of 2D clustering with annual data and various pairs of features.

Effective comparison of the results is obtained for the given dataset by calculating the clustering validity indicators, taking into account the output vector obtained from the same clustering algorithm (kmedoids) and always using the same set of initial data (reference power and annual energy) for calculating the indicators. Table I shows the numerical outcomes. For each indicator, the best case is highlighted in bold, and the second best is reported in italic. The 3D case and

two 1D cases (with P or U as features) are never considered as the best solutions, as well as the 2D case with P and U as features. Even though there is no clear best case, the cases that use the annual energy as features (in 1D or 2D) emerge as most promising according to all the indicators.

TABLE I. CLUSTERING VALIDITY INDICATORS FROM THE CLUSTERING RESULTS EXECUTED WITH ANNUAL DATA OF THE YEAR 2020 (FOR EACH INDICATOR: BEST IN BOLD, SECOND BEST IN ITALIC)

Indicator	Case and features used						
	3D (P,E,U)	2D (P,E)	2D (P,U)	2D (E,U)	1D (P)	1D (E)	1D (U)
MIA	0.788	*0.786*	1.097	**0.780**	1.488	0.836	1.187
CDI	0.339	0.320	0.766	*0.308*	0.778	**0.272**	1.676
DBI	*1.195*	**0.909**	2.086	2.123	3.207	1.532	9.156
MDI	20.47	**8.90**	26.09	24.33	17.47	*9.94*	58.41
WCBCR	0.0115	0.0102	0.0586	*0.0094*	0.0606	**0.0074**	0.2810

C.3. Impact of the COVID-19 Pandemics on Energy Consumption and Clustering Validity Indicators

From the previous results, the consideration of annual values has provided useful hints. However, during the year 2020 the effects of the COVID-19 pandemics had specific consequences on the users, especially from March to May 2020, within the main periods of lockdown (Fig. 9a). The main effects appear on the reduction of the number of users connected in these months, and in the reduction of the monthly energy, with the lowest values in April 2020.

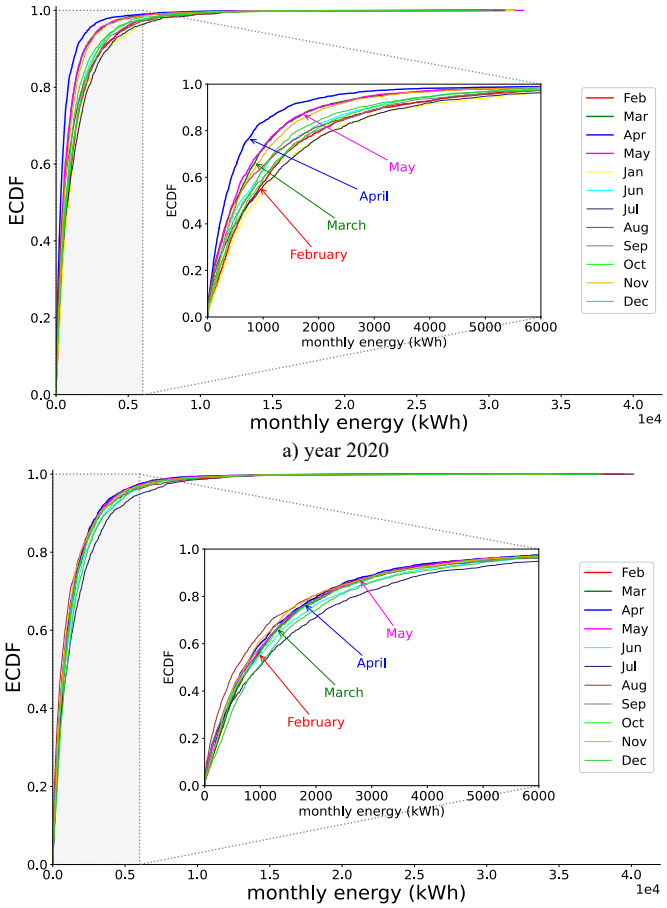
The same analysis previously carried out on the full year 2020 has been repeated in two different situations:

a) Using the data for the year 2020, by excluding the months March, April and May 2020. In this case, the correlation coefficients between the pairs of features ($\rho_{P,E}$ = 0.731, $\rho_{P,U}$ = 0.181, and $\rho_{E,U}$ = 0.533) are slightly higher with respect to the case with full annual data.

b) Taking the data of the users in the year 2019, considering the users also active in the year 2020. The correlation coefficients calculated for the year 2019 between the pairs of features are $\rho_{P,E}$ = 0.713 (relatively similar to the year 2020), while $\rho_{P,U}$ = 0.229 and $\rho_{E,U}$ = 0.674 are over 20% higher than in the year 2020. This is mainly due to the more uniform ECDFs of the monthly energy consumption in the months of 2019, as indicated in the ECDFs shown in Fig. 9b.

Fig. 9 shows the empirical CDF (ECDF) referring to the users that were connected for all the year 2020, from which the energy reduction in the lockdown period is clearly visible. In comparison, the ECDF referring to the monthly energy for the year 2019 is apparently more regular, as in 2019 there were no exceptional situations[1].

Table II shows the clustering validity indicators obtained from the analysis of the 9-month data in 2020. Even though a direct numerical comparison with the outcomes shown in Table I is not possible because the dataset used is not the same, the positions of the best cases for each indicator are unchanged, and also the positions of the second-best cases remain almost the same. Table III shows the clustering validity indicators obtained from the analysis of the annual data in 2019. The results confirm that the positions of the best cases or second-best cases for each indicator remain the same as in Table II. From these results, the general remark on the importance of considering the energy feature remains valid. The combination of the energy feature with the reference power or the utilisation is effective, with no specific need of taking the three features together.



a) year 2020

b) year 2019

Fig. 9. Empirical CDFs of the monthly energy for the years 2020 and 2019 with highlights on relevant months for the analysis.

TABLE II. CLUSTERING VALIDITY INDICATORS FROM THE CLUSTERING RESULTS EXECUTED WITH 9-MONTHS DATA (EXCLUDING MARCH-MAY 2020) (FOR EACH INDICATOR: BEST IN BOLD, SECOND BEST IN ITALIC)

Indicator	Case and features used						
	3D (P,E,U)	2D (P,E)	2D (P,U)	2D (E,U)	1D (P)	1D (E)	1D (U)
MIA	*0.760*	0.767	1.047	**0.755**	1.429	0.807	1.039
CDI	0.334	0.332	0.724	*0.310*	0.732	**0.302**	1.654
DBI	*1.178*	**0.890**	2.022	2.100	2.762	1.598	9.160
MDI	19.19	**9.58**	29.97	23.87	33.16	*10.69*	54.83
WCBCR	0.0111	0.0110	0.0525	*0.0096*	0.0536	**0.0091**	0.2734

TABLE III. CLUSTERING VALIDITY INDICATORS FROM THE CLUSTERING RESULTS EXECUTED WITH ANNUAL DATA OF THE YEAR 2019 (FOR EACH INDICATOR: BEST IN BOLD, SECOND BEST IN ITALIC)

Indicator	Case and features used						
	3D (P,E,U)	2D (P,E)	2D (P,U)	2D (E,U)	1D (P)	1D (E)	1D (U)
MIA	0.222	*0.213*	0.832	**0.184**	0.896	0.180	0.437
CDI	0.134	0.129	0.831	**0.105**	0.876	*0.106*	1.241
DBI	*1.093*	**0.847**	1.621	2.062	3.011	1.360	7.233
MDI	5.89	*3.53*	27.55	5.43	23.23	**1.91**	21.73
WCBCR	0.0018	0.0017	0.0690	*0.0011*	0.0770	**0.0011**	0.1540

[1] Successive years have not been considered, because in the year 2021 some effects of the pandemics continued, and in the year 2022 other external unexpected events resulted in considerable changes in energy prices and energy consumption.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2024.3379302

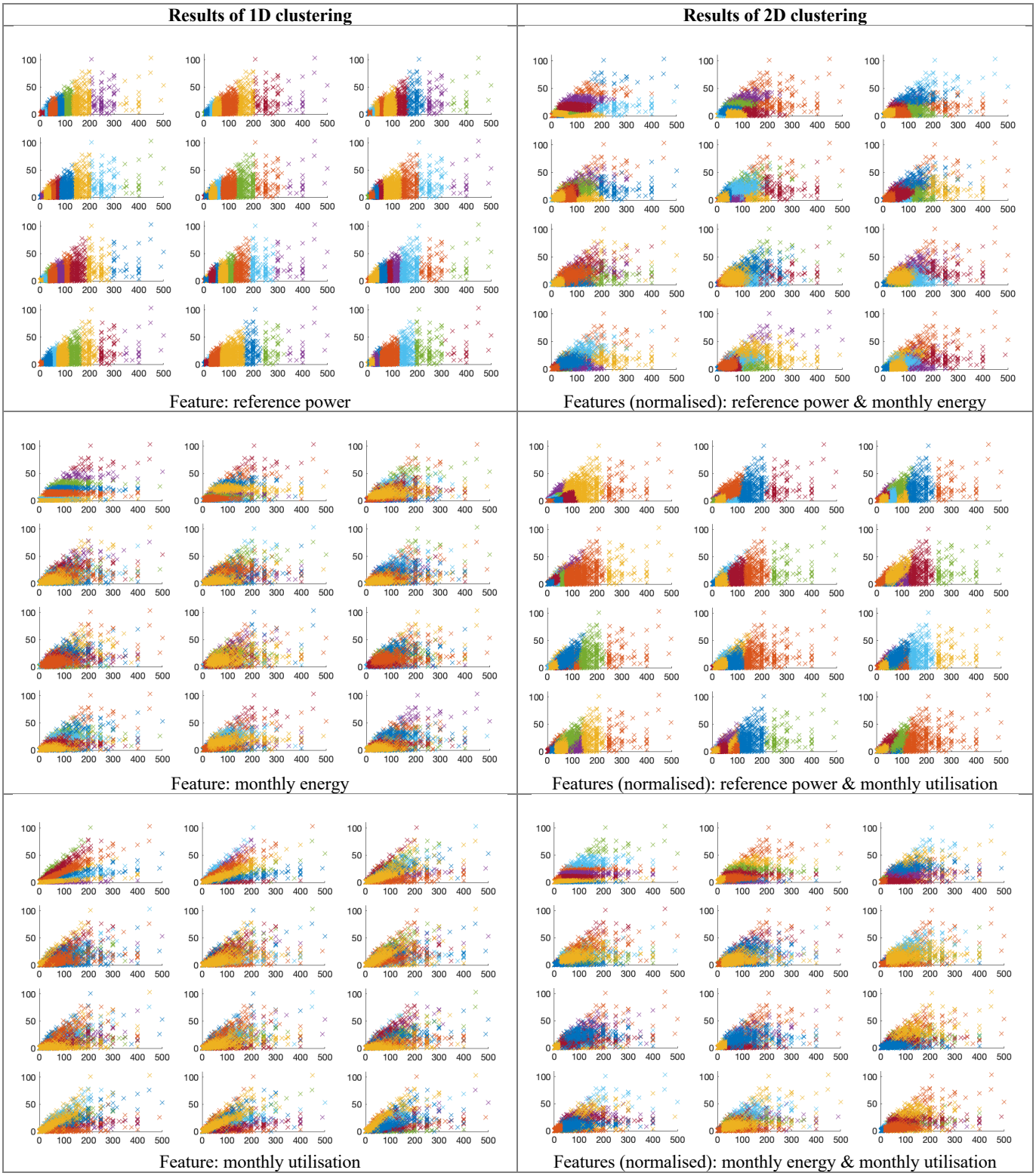


Fig. 10. Comparison among the clustering results for the 12 months (K = 10 clusters) on the same reference plot. For all plots: horizontal axis: power [kW]; vertical axis: monthly energy [MWh]. On each plot, the months are in sequential order by rows.

C.4. Analysis with Monthly Data

The reference power and the monthly energy are available for each month of the year 2020. The monthly utilisation has then been computed. The kmedoids clustering algorithm has been applied to the *separate monthly data.* Fig. 10 shows the solutions obtained in the individual months by using the various cases with 1D and 2D features. In the different months, there is no correspondence among the clusters and the colours with which the clusters are represented, as the cluster composition is not the same. The cluster partitioning differs in the various

months, even though the grouping of many users with relatively high power tends to remain similar, as well as for users with relatively low power. Fig. 11 shows the solutions obtained in the 12 months by using the 3D features.

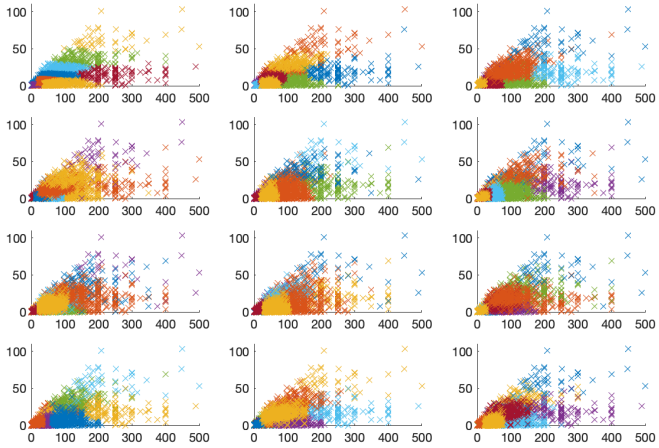


Fig. 11. Monthly clustering results by using the 3D features as inputs. Horizontal axis: reference power [kW]. Vertical axis: monthly energy [MWh]. The months are ordered by rows.

Based on the monthly data, the spectral clustering-based *consensus clustering* procedure has been executed to form the internal categories based on the clustering results. The consensus clustering procedure aggregates into a single output the results of the individual clustering algorithms executed for each month with a given set of features.

Table IV shows the clustering validity indicators obtained from the consensus clustering algorithm. The orders of magnitude are similar to those resulting in Table I for annual data (or better for the *MDI* index). In this respect, it has to be noted that the indicators have been calculated by taking the clustering results from consensus clustering and the annual data as references (to guarantee comparability among the outcomes).

The key point is that the synthesis of the monthly data given by consensus clustering takes into account the variability of the monthly data and provides meaningful results which depend on possible variable grouping of the users in different months. From the results indicated in Table IV, the basic remarks concerning the nature of the features remain similar as in the previous cases, namely, the results obtained when the energy feature appears (alone or in combination with other features) are more relevant than the results shown in the cases with reference power and utilisation.

TABLE IV. CLUSTERING VALIDITY INDICATORS FROM THE SPECTRAL CLUSTERING-BASED CONSENSUS CLUSTERING RESULTS WITH MONTHLY DATA IN THE YEAR 2020 (FOR EACH INDICATOR: BEST IN BOLD, SECOND BEST IN ITALIC)

Indicator	Case and features used						
	3D (*P*,*E*,*U*)	2D (*P*,*E*)	2D (*P*,*U*)	2D (*E*,*U*)	1D (*P*)	1D (*E*)	1D (*U*)
MIA	0.931	*0.918*	1.133	**0.894**	1.529	0.992	1.579
CDI	0.331	*0.308*	0.790	**0.289**	0.819	0.324	1.846
DBI	*1.48*	**1.18**	2.50	3.33	8.41	1.97	17.68
MDI	5.36	*2.85*	15.52	27.27	374.07	**2.66**	39.68
WCBCR	0.011	*0.010*	0.062	**0.008**	0.067	0.011	0.341

The comparison with the results obtained for the year 2019 (Table V) confirms the general effectiveness of using the energy feature, with slight differences referring to the prevailing effectiveness of using the (*P*,*E*) pair of features for most indicators and the presence of the three features (*P*,*E*,*U*) as the best or second best cases for all indicators.

TABLE V. CLUSTERING VALIDITY INDICATORS FROM THE SPECTRAL CLUSTERING-BASED CONSENSUS CLUSTERING RESULTS WITH MONTHLY DATA IN THE YEAR 2019 (FOR EACH INDICATOR: BEST IN BOLD, SECOND BEST IN ITALIC)

Indicator	Case and features used						
	3D (*P*,*E*,*U*)	2D (*P*,*E*)	2D (*P*,*U*)	2D (*E*,*U*)	1D (*P*)	1D (*E*)	1D (*U*)
MIA	*0.218*	**0.212**	0.836	0.223	0.896	0.246	0.457
CDI	*0.133*	**0.128**	0.840	0.135	0.879	0.158	0.991
DBI	*1.05*	**0.85**	1.47	1.49	3.08	2.55	8.26
MDI	**2.67**	*3.19*	25.86	4.31	23.23	4.99	16.65
WCBCR	*0.0018*	**0.0016**	0.0705	*0.0018*	0.0772	0.0025	0.0981

IV. CONCLUSIONS

This paper has introduced the original definition of internal categories referring to each macro-category, showing its application to low-voltage three-phase electricity users. By constructing the internal categories, the DSO can obtain more information on the partitioning of the overall consumption of the users connected to a portion of its network, by exploiting the limited data available about the users (reference power and annual or monthly energy, from which the utilisation can be determined as a further feature). The overall scheme that leads to DSO-based load profiling has been upgraded accordingly.

With the introduction of the internal categories, there is the possibility to group the consumers based on their power and energy characteristics, forming more refined classes that can be sent to the successive stratified sampling procedure aimed at identifying how many customers should be monitored within each class with the same statistical significance. Establishing the limits of the classes is crucial for low-voltage three-phase electricity users, due to the variety of their reference power and energy consumption values.

From the analysis with annual data and the calculation of clustering validity indicators, the annual energy has been found to be the most important feature. The combination of the annual energy feature with the reference power or annual utilisation to form 2D data has proven effective, without the specific need to use the three features together. This result has been confirmed by the analysis with monthly data, also combined through a specific consensus clustering procedure.

The analysis has been carried out with data gathered in two years (2019 and 2020). In particular, in the year 2020, specific situations occurred in some months due to the effects of the COVID-19 pandemic, during the lockdown period imposed in the region where the DSO is located. The most suitable features (energy, also used with another feature in 2D data) that emerged from the study with the annual data for the years 2019 and 2020 have been confirmed also after removing the lockdown months occurred in 2020 from the determination of the internal categories.

In the months affected by the lockdown due to pandemic, separate clustering has been executed on each month, and the consensus clustering procedure has been applied for aggregating the results into a single final set of clusters. The consensus clustering results have confirmed once again that the cases with 2D features in which the monthly energy is included

are effective to form the internal categories. In the comparisons that considered the period of lockdown due to COVID-19, the users taken into account have been the ones that did not close their activity during the year 2020. These users could have had reductions in their energy consumption. However, no remarkable impact has been found on the relevance of the features to be used for creating the internal categories. The analysis did not take into account the users that closed their activity (permanently or temporarily), suffering the most critical consequences of the pandemic. The users with temporary activity may be added to the internal categories formed by determining the internal category with the minimum distance from the medoids of the formed clusters based on the data of these users in non-lockdown periods.

Following the recent trends to apply more data-driven approaches based on the numerous data made available by 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 necessary if the electrical behaviour of one or more users changes based on the information available from a continuous stream of data. In this regard, the creation of the internal categories can be repeated at any time horizon based on the user-defined features.

The results have shown that there is no single solution to the formation of the internal categories when the set of features changes. From the explored cases, data normalisation and the use of annual or monthly data within a consensus clustering strategy have relevant effects on the clustering results. The specific effects of introducing the three types of features have been shown. For example, the objective of obtaining a grouping based primarily on the reference power partitioning can be approached by including the reference power feature. However, more significant results on the overall behaviour have been obtained with the 2D cases that include the energy feature.

The ongoing activity is aimed at determining the minimum number of users to monitor within each internal category to carry out the subsequent load profiling steps. In this regard, the use of multi-dimensional features for the internal categories requires the adaptation of the sampling procedures, moving from classical procedures that consider 1D data to specific procedures capable to perform 2D sampling.

V. Appendix

A. Evolution of the COVID-19 Pandemic in the Relevant Territory

The period of analysis contains data referring to the year 2020. In some months of 2020, the electricity usage by different categories of users was heavily affected by the COVID-19 pandemic. The assessment of the data referring to the period of the COVID-19 pandemic is linked to the sequence of actions and restrictions applied in the jurisdiction of the Piedmont region (where the users under analysis are located).

Some relevant historical notes follow. The first case of COVID-19 was identified in Italy on 21 February 2020. A few days later, on 11 March 2020, the World Health Organisation declared COVID-19 as a worldwide pandemic [37]. Like many other countries, Italy has imposed quarantines and restrictions to prevent the further spread of the pandemic and avoid the collapse of its healthcare system, put under pressure by the increasing number of people hospitalised in intensive care.

On 1 March 2020, the Italian Council of Ministers approved a decree that divided Italy into areas characterised by different levels of pandemic spread and severity of the restrictions adopted (red, orange, and yellow zones). In Piedmont, initially less severe restrictions were chosen; safety and prevention measures were advertised in public places, and special sanitisations were performed on public transport. On 4 March 2020 the situation worsened, and the Italian Government began to adopt stricter lockdowns across the entire country. First, the shutdown of all schools and universities nationwide was imposed for two weeks, also banning public attendance at all sporting events. Then, on 8 March 2020, the Government locked down the Lombardy region and 14 other provinces in Northern Italy, including Piedmont. Any movement in and out of the areas was prohibited, except for emergencies or "proven working needs". With the same decree, the Government also established restrictions for shopping centres and commercial activities. On 9 March 2020, the Government announced that all sporting events in Italy would be cancelled until at least 3 April 2020, with the only exception of international football competitions. On the same day, all the measures previously applied only in the “red zones" were extended to the entire country. On 11 March 2020, all commercial and retail businesses were closed, the only exceptions being those providing essential services. On 20 March 2020, the Ministry of Health ordered tighter regulations on free movement, by banning "any movement towards a residence other than the main residence", including holiday homes, during weekends and holidays. On 21 March 2020, further restrictions were announced as part of the nationwide lockdown, to stop all non-essential productions, industries, and businesses in Italy.

Thanks to these strong actions, the number of infections has decreased. Starting from 18 May 2020, most businesses were able to reopen, and free movement was guaranteed to all citizens within their region; travelling between regions was still banned for non-essential reasons. Swimming pools and gyms could also reopen on 25 May 2020, followed by theatres and cinemas on 15 June 2020. On 3 June 2020, free movement was restored throughout the national territory, *de facto* ending the lockdown phase that began in March 2020.

Since October 2020, following the growth in infections after the summer, further restrictions have been imposed to specific activities. These restrictions further affected the consumption in the last three months of 2020 and beyond.

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] 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.

[3] 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.

[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] 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.

[6] S. Pelekis, A.Pipergias, E. Karakolis, S. Mouzakitis, F. Santori, M. Ghoreishi, and D. Askounis, "Targeted demand response for flexible energy communities using clustering techniques," *Sustainable Energy, Grids and Networks*, vol. 36, ref. 101134, 2023.

[7] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications", *IEEE Trans. Smart Grid*, vol. 7 (5), pp. 2437–2447, 2016.

[8] D. Qiu, Y. Wang, J. Wang, C. Jiang, and G. Strbac, "Personalized retail pricing design for smart metering consumers in electricity market," *Applied Energy*, vol. 348, ref. 121545, 2023.

[9] M.B. Rasheed and M.D.R. Moreno, "Minimizing pricing policies based on user load profiles and residential demand responses in smart grids," *Applied Energy*, vol. 310, ref. 118492, 2022.

[10] A. Sumaiti, S.R. Konda, L. Panwar, V. Gupta, R. Kumar, and B.K. Panigrahi, "Aggregated Demand Response Scheduling in Competitive Market Considering Load Behavior Through Fuzzy Intelligence," *IEEE Trans. Industry Appl.*, vol. 56 (4), pp. 4236-4247, 2020.

[11] J.N. Fidalgo, M.A. Matos, and L. Ribeiro, "A new clustering algorithm for load profiling based on billing data," *Electric Power Systems Research*, vol. 82, no. 1, pp. 27–33, 2012.

[12] G. Chicco, D. Bonansinga and P. Colella, "Categorisation of Low-Voltage Three-Phase Electricity Users," *Proc. 2022 International Conference on Smart Energy Systems and Technologies (SEST)*, Eindhoven, Netherlands, pp. 1-6, 2022.

[13] H. Zhong, Z. Tan, Y. He, L. Xie, and C. Kang, "Implications of COVID-19 for the Electricity Industry: A Comprehensive Review," *CSEE Journal of Power and Energy Systems*, vol. 6 (3), pp. 489–495, 2020.

[14] S. García, A. Parejo, E. Personal, J.I. Guerrero, F. Biscarri, and C. Léon, "A retrospective analysis of the impact of the COVID-19 restrictions on energy consumption at a disaggregated level," *Applied Energy*, vol. 287, ref. 116547, 2021.

[15] M. Ferrando, A. Banfi, and F. Causone, "Changes in energy use profiles derived from electricity smart meter readings of residential buildings in Milan before, during and after the COVID-19 main lockdown," *Sustainable Cities and Society*, vol. 99, art. 104876, 2023.

[16] C. Si, S. Xu, C. Wan, D. Chen, W. Cui, and J. Zhao, "Electric Load Clustering in Smart Grid: Methodologies, Applications, and Future Trends," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 2, pp. 237–252, 2021.

[17] C.L. Athanasiadis, T.A. Papadopoulos, G.C. Kryonidis, and D.I. Doukas, "A review of distribution network applications based on smart meter data analytics," *Renewable and Sustainable Energy Reviews*, vol. 191, art. 114151, 2024.

[18] 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.

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

[20] 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.

[21] 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.

[22] C. Ribeiro, T. Pinto, Z. Vale, and J. Baptista, "Customized normalization clustering methodology for consumers with heterogeneous characteristics," *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 7, no. 2, pp. 53–69, 2018.

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

[24] 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.

[25] 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.

[26] M.R. Anderberg, *Cluster Analysis for Application*, Academic Press, New York, 1973.

[27] L.G.B. Ruiz, M.C. Pegalajar, R. Arcucci, and M. Molina-Solana, "A time-series clustering methodology for knowledge extraction in energy consumption data," *Expert Systems with Applications*, vol. 160, 113731, 2020.

[28] Mathworks, k-medoids clustering, [online] https://it.mathworks.com/help/stats/kmedoids.html (accessed 12 January 2024).

[29] 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.

[30] 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.

[31] D.L. Davies, and D.W. Bouldin, "A cluster Separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAM-1 (2), pp. 224–227, 1979.

[32] J.C. Dunn, "Well separated clusters and optimal fuzzy partitions," *J Cybernetics*, vol. 4, pp. 95–204, 1974.

[33] G.J. Tsekouras, N.D. Hatziargyriou, and E.N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 22 (3), pp. 1120–1128, 2007.

[34] A. Radovanović, X. Ye, J.V. Milanović, N. Milosavljević, and R. Storchi, "Application of the k-medoids Partitioning Algorithm for Clustering of Time Series Data," *Proc. 2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, October 26-28, 2020.

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

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

[37] Coronavirus disease (COVID-19) pandemic [Online] Available: https://www.who.int/europe/emergencies/situations/covid-19. [Accessed: 12 January 2024].

Gianfranco Chicco (M'98, SM'08, F'18) holds a Ph.D. in Electrotechnics Engineering and is a Full Professor of Electrical Energy Systems at Politecnico di Torino, Italy. He is the 2023-2024 Chair of the IEEE R8 Italy Section. He is the Editor-in-Chief of Sustainable Energy Grids and Networks. He was the Conference Chair of WESC 2006, IEEE PES ISGT Europe 2017, UPEC 2020 and IEEE Eurocon 2023. His research topics include Power System Analysis, Distribution System Analysis and Optimization, Electrical Load Management, Energy Efficiency and Environmental Impact of Multi-Energy Systems, Data Analytics Applied to Power and Energy Systems, and Power Quality. He is a member of the Italian Association AEIT.

Daniele Bonansinga received the M.S. degree in Electrical Engineering at Politecnico di Torino, Italy, in 2022. He is now an electrical engineer and working in the field of Smart City development at the multi-utility company IREN Torino. His work focuses on finding solutions to make the city more secure, sustainable, efficient, and innovative, a city capable of ensuring a high quality of life for its citizens.

Pietro Colella (M'15) holds a Ph.D. in Electrical Engineering and is an assistant professor at Politecnico di Torino, Italy. His main interests are power system analysis with traditional and advanced machine learning techniques, traction electrification systems, electricity markets, and electrical safety. He has been and is the principal investigator of several Industrial and National Research Projects in these fields. He is the author of more than 20 articles in scientific journals. He is a member of the Italian Association AEIT.

Lorenzo Solida (GSM'23) received the M.Sc. degree in Electrical Engineering from Politecnico di Torino, Italy, in 2019. From November 2020, he is a Ph.D. student in Electrical, Electronics, and Communications Engineering at the Energy Department (DENERG) of Politecnico di Torino. His research activity is focused on the impact and optimal allocation of renewable energy sources in the electrical power systems. Since 2020 he is a member of the Italian Society of Engineers and a member of the Italian Association AEIT.