Load pattern classification for tariff purposes is typically performed on aggregate residential load data, or on individual non-residential load data. Residential consumers are generally not classified as individual entities, for the following reasons:

- the consumption pattern of individual residential loads varies in function of the number of persons composing the family, as well as their activity, age and lifestyle [3][4]; the characterization of residential customers by taking into account the expected load pattern of each single customer would require performing a detailed statistical analysis based on the several factors affecting the energy use in a family [5]-[7]; however, the variation of the individual residential customer load at each hour of the day, mainly conditioned on the occasional use of a few facilities with relatively large power consumption (e.g., washing machine, electrical oven, and so forth) [8], is so large to make the use of statistical data impractical;
- the electrical distribution system lines starting from the MV/LV substation do not feed the residential loads directly, but each distribution system feeder supplies an aggregated load; in the presence of a significant number (e.g., a few dozens or more) of residential customers, the diversity of energy use for each customer makes the aggregated load pattern smoother with respect to the one of the individual residential load; the time evolution of the aggregated load pattern can be predicted to a relatively good extent [7].

Conceptually, electricity customer classification should follow the rules of segmentation referred to the commercial types of activity, as established for instance by the national institutes of statistics. However, there is a great diversity among the load patterns of the customers belonging to the same type of activity or associated to the same commercial code [9][10]. As such, customer classifications based on the type of activity and on commercial codes are not efficient for representing the specific aspects of the electricity consumption. The distinction should be limited to some macro-categories (e.g., residential, industrial, commercial, or other specific categories such as lighting and traction). Identification of some “external” features can be useful to make a preliminary customer partitioning into macro-categories.

1. INTRODUCTION

In most restructured electricity markets, distribution and supply services have been unbundled. Electricity suppliers are now operating within a competitive environment, with some degrees of freedom in formulating the tariff offers [1][2], provided that their offers meet the requirements set by regulatory authorities in the form of price or revenue caps.

Electrical load pattern classification is carried out with the main objective of identifying a suitable set of customer classes on the basis of the shape of the electrical load patterns. For the customer classes formed, the supplier can then formulate specific tariff offers.

Abstract: In the current structure of the electricity business, distribution and supply services have been unbundled in many jurisdictions. As a consequence of unbundling, electricity supply to customers is now provided on a competitive basis. In this context, the electricity suppliers need to get accurate information on the actual behaviour of the electricity customers, for the purpose of setting up effective commercial offers. Grouping the electrical load patterns on the basis of information on their activity or commercial codes has proven to be ineffective, since very different load patterns would result in the same group. Customer classification on the basis of consumption pattern similarity is likely to provide more effective results. In order to establish customer grouping based on similarity aspects, various clustering techniques have been tested on electrical load pattern data. This paper provides an overview of these techniques, included in a more general scheme for analyzing electrical demand data. The various stages of the customer classification procedure include the definition of the information to be gathered on the field, the selection of the features to be used to run the clustering methods, the use of clustering methods with assessment of their effectiveness through the calculation of appropriate clustering validity indicators, and the formation of the final load profiles representing a relatively limited number of final customer classes. The characteristics of these stages are illustrated and discussed, providing links to relevant literature references.

Keywords: Classification, Clustering, Electrical consumer, Electrical demand, Customer categorization, Load pattern, Load profile, Validity indicator.
Possible external features are the rated values of electrical quantities, the type of activity and other information such as supply voltage level, annual active and reactive energy (maximum, minimum, average value and standard deviation), utilization level (defined as the energy consumption to rated power ratio), and power factor. Moreover, it is possible to build separate models for weather-dependent loads. Using macro-categories, the number of load patterns to be handled together for each macro-category by the classification methods would be reduced.

Dedicated research has been developed in the last decade to study classification mechanisms based on the shape of the load patterns, useful for the formulation of tariff options dedicated to each load class. Furthermore, once identified, each class can be represented for tariff purposes through its synthetic load profile, on the basis of which the supplier can make its evaluations and the authorities can set up appropriate regulation.

Starting from an initial pre-defined number of macro-classes (for instance, residential, industrial, commercial and others), and from the identification of time periods with different consumption characteristics (e.g., weekdays and weekend days, defined for different periods of the year in order to take into account seasonality issues), research on load pattern classification has been carried out to formulate suitable algorithms able to make sound grouping of the customers belonging to the same macro-class in a given time period, using load pattern shape information gathered from results measured on the field. More specifically, the electrical load pattern classification approach starts from the assumption that it is possible to measure the load pattern of any customer belonging to the same macro-category, for a given duration of observation. Practically, on-site measurements have to be performed for a time period long enough to get a sufficient amount of data for customer classification. With the present diffusion of metering facilities (although suitable technologies are available), generally it is still not possible to perform measurements on every customer in every jurisdiction, even though the installation of smart meters is in progress in many countries. As such, load pattern classification can be performed by creating the customer classes by monitoring a limited number of customers. Then, a suitable mechanism for associating the other customers to the classes formed has to be identified. The minimum number of customers to be subject to load pattern measurement for the various macro-classes can be determined by using statistical techniques such as the stratified sampling approach [11].

The remainder of this paper addresses the various steps of the electrical load pattern classification process, according to the general scheme outlined in Figure 1 [12][13].

## 2. DATA GATHERING AND PROCESSING

Let us consider a set of $M$ customers belonging to the same macro-category, to be classified into a meaningful number of customer classes. Relevant data are referred to comparable periods in terms of type of day (weekday/weekend) and season [12][14]. Let us refer to the context characterizing each of these periods as a loading condition.

Typically, the monitored data are organised to represent the customer’s consumption by means of a daily load pattern. The duration of the monitoring period has to be long enough to guarantee availability of a sufficient amount of data. Hence, the corresponding monitoring period should be at least two-three weeks in the same loading condition.

The sampling rate depends on the characteristics of the monitoring equipment used to collect data (for a given monitoring period the rate limit can depend on data storage capability). The time intervals of interest for data representation are typically 1 minute, 15 minutes or one hour. The corresponding number of samples characterizing each daily load pattern is $H = 1440$, $H = 96$, and $H = 24$, respectively. One-minute sampling can be used to gather a sufficient number of points to compute 15-minute data by smoothing the effect of the discretization step of the measurement system [15]. Generally, even though faster measurements are done, the stored data refer to 15-minute time intervals [10][16]-[18].

![Figure 1. The load pattern classification procedure.](image)
Data accuracy depends on the characteristics of the monitoring equipment. In order to improve the accuracy of the information gathered, sometimes it would be better to monitor data with a sampling rate higher than the one corresponding to the time interval of interest (e.g., each minute), thus calculating the data related to the time interval of interest (e.g., 15 minutes) by averaging the single data monitored inside each time interval. Bad data detection and elimination is performed in such a way to ensure that the load patterns used for customer classification correspond to normal operating conditions. For this purpose, load data leading to uncommon situations are detected and eliminated. Practically, uncommon situations may occur depending on anomalous days (e.g., bank holidays occurring at weekdays), expected events (e.g., maintenance) or unexpected events (e.g., failures, strikes, ...). The effects of failures or abnormal conditions may be detected by identifying the time intervals at which the average RMS voltage is outside the acceptable range (90%–110% of the rated voltage). Moreover, a dedicated procedure for de-noising by wavelet multiresolution analysis is presented in [19]. The bad data detected are eliminated from the analysis and the number of useful data for a given time interval is correspondingly reduced.

In order to represent the customer information, the average load pattern is determined by computing the average value of the useful load patterns gathered at each time interval. For instance, if measurements have been performed for two weeks in the loading condition corresponding to the spring season, measuring 10 weekend days, the average load pattern for the representative weekday is calculated by averaging instant-by-instant the 10 weekday load patterns (assuming there is no anomalous day). If a bad data appears, for instance at hour 9 am for one of the days, this bad data is excluded from the averaging, and the remaining 9 points are used to determine the average value at hour 9 am to be included in the average daily load pattern.

The averaging process allows for smoothing the average load pattern curve, thus reducing the relevance of possible power values obtained in “normal” operating conditions but largely outside the average load.

For a given loading condition, the information concerning each customer, to be used for classification purposes, is thus given in such a way to get load patterns comparable in terms of their shape. The information stored contains:

- the reference power [kW], defined as the peak value of the average load pattern;
- the normalised representative load pattern (RLP), computed by dividing the average load pattern by its reference power.

The effect of this definition is that the reference power does not correspond to the true peak power reached by the load pattern in the period of observation, because of averaging multiple points at corresponding time instants of the different measured days. However, this fact can be seen in a positive way, since non-regular peaks that could occur during the measurements come out to have a limited impact on the RLP. The normalization aspect is extended in [20] to include also the minimum value of the load pattern, in such a way that all RLPs formed have a null minimum value and a unity maximum value.

3. PRE-CLUSTERING PHASE

3.1. Definition of the features to be used for classification purposes

Feature selection concerns the identification of the type of data to be used for performing the evaluations referred to load classification.

The initial data are the RLPs built from the measured time-domain data. The time-domain RLPs can be used directly, or can be processed to obtain other features representing the customers. Using time-domain data, the load patterns are defined with an arbitrary number of average power values, depending on the meter resolution. For a given loading condition, a simple way to define the features of the \( m^{th} \) representative load pattern, for \( m = 1, \ldots, M \), is to consider all or a part of the normalised power values obtained from the measurements in the time domain. In this way, a set of \( H \) directly determined shape features is readily available, without performing any load pattern post-processing. In the time domain, an analysis could be made for grouping together a number of successive time intervals. For instance, the RLP data corresponding to each 15 minutes could be grouped together in such a way to identify a reduced number of time intervals, composed of night hours (from 0 am to 6 am), sunrise hours (from 6 am to 8 am), morning hours (from 8 am to 12 am), lunchtime hours (from 12 am to 2 pm), and so forth [12]. Let us denote the set of RLPs as \( \mathbf{X} = \{x^{(m)}, m = 1, \ldots, M\} \), whose \( m^{th} \) component is represented by the vector \( x^{(m)} = \{x^{(m)}_h, h = 1, \ldots, H\} \).

An alternative to the use of time-domain data is the definition of suitable indirectly determined shape features, requiring post-processing of the time-domain data for their definition. While determining the shape features, an interesting possibility refers to reducing the number of data to be stored for each customer and to be sent to the classification tools. Among the shape features, it is possible to define a set of shape factors, modelling specific aspects of the customer consumption “signature”. The shape factors are defined for each customer on the basis of the representative load diagram in a given loading condition. Examples are the dimensionless ratios used in [2], [10], [12] and [21], related to average to maximum power ratios, or ratios between average power at different portions of the day (daylight period, night period, or the entire day). Other types of indirectly determined shape features are those identified in the frequency domain, such as the harmonics-based coefficients presented in [12] and [22], the Fourier series coefficients used in [23], and the
coefficients derived from the wavelet transform exploited in [24]. Furthermore, data size reduction can be performed by using projection methods, such as the Principal Component Analysis (PCA), Curvilinear Component Analysis (CCA) and Sammon Map exploited in [25], or the Canonical Variate Analysis (CVA) used in [26].

3.2. Load pattern data processing to build the input data set

The input data for a given loading condition can be conveniently set up in the form of a matrix, for instance with \( M \) rows (for the \( M \) customers) and \( H \) columns (for the \( H \) features), with an additional column vector of \( M \) components containing the reference powers for every customer. The clustering techniques can use the RLPs to form the groups. The reference powers can be either ignored during the group formation (thus giving the same conceptual importance to all load patterns regardless of the corresponding actual power), or can be more conveniently exploited as weighting factors in the calculation of the centroids as weighted sums of the load patterns contained in the group. In this way, the centroid will assume a more meaningful role. After creating the centroid, it is important to associate to the centroid the appropriate reference power, given by the sum of the reference powers of the load patterns belonging to the group represented by the centroid. Once created, the centroid will generally have a maximum value lower than unity.

4. CLUSTERING PHASE

4.1. Clustering techniques

On the basis of the features defined, the core of the classification procedure is the adoption of an effective classification technique. Clustering techniques [27]-[29] are generally used to perform this task. In particular, it is possible to identify [23] unsupervised learning-based techniques, such as the Kohonen’s self organizing map (SOM), supervised learning-based techniques, such as the ones adopting multilayer perceptron or Elman neural networks, or vector quantization, fuzzy logic-based techniques, statistical techniques such as k-means (KM) and multivariate analysis, and hybrid techniques such as probability neural networks (PNN) and fuzzy k-means (FKM). Further techniques have been recently defined by following the concept of entropy borrowed by information theory, or adapting classification techniques used in other domains, such as follow the leader (FDL) and support vector clustering (SVC). A summary of the techniques used in various literature papers, with indication of relevant references, is shown in Table 1.

On the application side, the clustering techniques differ according to the principle used in their definition, but can be discussed on the basis of the requirements for their usage. A first aspect is the possibility of setting up the final number of clusters the user intends to obtain.

This possibility can be of interest for the supplier or for the regulating authority, since in their perspective the number of final consumer classes cannot be too high, in order to set up a relatively small number of tariff options, whose contents and differences have to be readily understandable by the consumers. In this respect, the different methods behave as follows:

a) Agglomerative techniques such as the hierarchical clustering [20][21][27][30] can be easily adopted to produce a given number of clusters. In fact, the hierarchical clustering procedure starts with a number of classes equal to the number of RLPs, and proceeds by adding one load pattern at a time to the “closest” existing class (according to specific linkage criteria) up to reaching the desired number of clusters. However, the hierarchical clustering includes no mechanism for improving the cluster formation by reassigning the load patterns to the clusters already formed, and the performance obtained from its variants (with different linkage criteria [31]), evaluated by using appropriate clustering validity indicators (see section 4.2), are relatively different. The agglomeration principle is also used in the approach illustrated in [32], based on information theory principles [33] and adopting an effective non-linear metric exploiting Renyi entropy concepts [34][35] in the development of the clustering algorithm.

<table>
<thead>
<tr>
<th>method</th>
<th>references</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5.0</td>
<td>[10] (2005)</td>
</tr>
<tr>
<td>Entropy-based (Renyi)</td>
<td>[32] (2010)</td>
</tr>
<tr>
<td>Fuzzy and ARIMA</td>
<td>[40] (2005)</td>
</tr>
<tr>
<td>Iterative Refinement Clustering (IRC)</td>
<td>[41] (2005)</td>
</tr>
<tr>
<td>Multivariate statistics (MANOVA)</td>
<td>[23] (2006)</td>
</tr>
<tr>
<td>Support vector clustering (SVC)</td>
<td>[42] (2009)</td>
</tr>
<tr>
<td>Weighted Evidence Accumulation Clustering (WEACS)</td>
<td>[21] (2007)</td>
</tr>
</tbody>
</table>
b) Other techniques such as k-means and fuzzy k-means [27][44], accepting the final number of clusters as input, in a few cases resulted in forming a lower number of clusters with respect to what required, due to the presence of empty clusters in the final grouping. However, since the procedure of the method is not deterministic, with internal steps depending on random number extractions, it is possible to find a suitable solution with the desired number of clusters by running the method again. Basic k-means and fuzzy concepts have been exploited to set up clustering algorithms applied to load pattern classification [19][20][36][45]. The Adaptive Vector Quantization (AVQ) method used in [20] is an unsupervised one-layer neural network that uses a competitive layer with a constant number of neurons. Customized versions using fuzzy principles [46] have been proposed, as in [39] and [47] by exploiting Min-Max neuro-fuzzy network [48] concepts. Moreover, a fuzzy inference model using fuzzy rules to identify the input data, as well as to create local regression models, is illustrated in [45], and a study showing the possibility of combining fuzzy clustering and Auto-Regressive Integrated Moving Average (ARIMA) statistical models is reported in [40].

c) Conversely, the follow the leader algorithm [2][49] does not require the definition of the number of clusters as input, but uses an internal distance threshold among the cluster centroids, whose variation produces different numbers of clusters in a deterministic way (that is, the number of clusters obtained with a given distance threshold is always the same for the same set of initial data). Since the follow the leader algorithm is relatively fast, it is possible to run the algorithm more times successively, with different distance thresholds, until the specified number of clusters is reached.

d) The IRC method [41] has been defined to merge the most interesting properties of the hierarchical clustering (working with a specified number of clusters) with the ones of the follow-the-leader (the presence of an iterative mechanism for reassigning the load patterns to the clusters already formed), and includes an explicit mechanism to avoid the formation of empty clusters.

e) The SOM [50] modifies the search space to represent the results on a bi-dimensional map [14][37][38], but does not generate directly the final clusters. Hence, a post-processing stage is needed to form the clusters, with arbitrary assumptions, so that different numbers of clusters can be formed starting from the same SOM outcomes, by using a specific technique to identify the final clusters (for instance, post-processing based on k-means is used in [37] and [10]). Likewise, SVC [51] requires a first stage in which the support vectors are formed, followed by a second stage in which the groups are formed for the desired number of clusters [42]. Furthermore, the statistical multivariate technique MANOVA is presented in [23], by highlighting its graphical representation capability, similar to the one of the SOM, that allows for simple and effective visualization of the clustering results.

f) Further applications have been performed in [43] by using Probabilistic Neural Networks (PNN), based on finding for each load pattern the class with maximum probability of being the right one, and in [21] by using a Weighted Evidence Accumulation Clustering (WEACS) approach.

An important aspect is that the clustering algorithm can be executed on the basis of different features. However, for the sake of comparison among the final results from different clustering techniques, the final RLP grouping has to be made on the basis of time-domain data, whatever feature has been used to run the clustering procedure. More specifically, regardless of the specific details of the clustering method, the only output needed from the clustering algorithm is the allocation of the RLPs to the clusters formed. This can be done by constructing a two-dimensional list, in which the first dimension contains the number of clusters 1,...,K, while the second dimension contains for each cluster the list of RLPs belonging to that cluster. Alternatively, it is possible to create and handle a unique vector containing M components, progressively updated during the clustering process, in which the mth component contains the number of the cluster to which the mth RLP is assigned.

4.2. Clustering validity indicators

Different clustering validity indicators have been defined in order to assess the effectiveness of the clustering methods. Most of these indicators are based on Euclidean distance metrics. For this purpose, different types of distance are needed. A comprehensive vector-based formulation of the distances is presented in [12] and [13]. Assuming that the clustering results originate the set of centroids \( C = \{e^{(k)}, k = 1,...,K\} \) and the corresponding groups of the RLPs, denoted as \( L^{(k)} \), each of which contains \( n^{(k)} \) RLPs, for \( k = 1,...,K \), it is possible to consider various distances. Applying the Euclidean distance rationale, the set of distances used includes the pattern-to-pattern distance, for instance \( d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \) between the \( i^{th} \) and \( j^{th} \) RLPs, the pattern-to-set distance, for instance \( d(\mathbf{x}^{(i)}, L^{(k)}) \) from the \( i^{th} \) RLP and the \( k^{th} \) clustered group, the average set-to-set distance, for instance \( d(L^{(i)}, L^{(j)}) \) between the \( i^{th} \) and \( j^{th} \) clustered groups, and the infra-set distance, for instance \( \hat{d}(L^{(k)}) \) referred to the \( k^{th} \) clustered group.

Starting from these definitions, some clustering validity indicators have been defined under the common rationale according to which, for each indicator, lower values represent better clustering validity. For this
purpose, the original definitions of some indicators have been modified. The set of clustering validity indicators used in various publications, referred to the formation of \( K \) clusters, is the following:

- Mean Index Adequacy (\( MIA(K) \)) [2]:
  \[
  MIA(K) = \frac{1}{K} \sum_{k=1}^{K} d^2(e_k, L_k)
  \]  
  (1)

- Clustering Dispersion Indicator (\( CDI(K) \)) [2]:
  \[
  CDI(K) = \frac{1}{d(C)} \sqrt{\frac{1}{K} \sum_{k=1}^{K} d^2(L_k)}
  \]  
  (2)

- Scatter Index (\( SI(K) \)) [52]:
  \[
  SI(K) = \left( \sum_{n=1}^{M} d^2(x_n, p) \right) \left( \sum_{k=1}^{K} d^2(e_k, p) \right)^{-1}
  \]  
  with pooled scatter \( p = \frac{1}{M} \sum_{n=1}^{M} x_n \).

- Variance Ratio Criterion (\( VRC(K) \)) [53]:
  \[
  VRC(K) = \frac{1}{M} \left( 1 + \frac{W}{K - 1} \left( 1 - \frac{W}{M - K} \right) \right)^{-1}
  \]  
  (4)
  where \( W = \frac{1}{K} \sum_{i=1}^{K} \left( n_i - 1 \right) \left( 1 - \frac{n_i}{M} \right) \).

- Davies-Bouldin Index (\( DBI(K) \)), considering the version of the index introduced in [54] constructed by using the Euclidean distances and for \( i, j = 1, \ldots, K \):
  \[
  DBI(K) = \frac{1}{K} \sum_{i \neq j} \max \left\{ \frac{d(X_i, p) + d(X_j, p)}{d(e_i, e_j)} \right\}
  \]  
  (5)

- Similarity Matrix Indicator (\( SMJ(K) \)) [12], for \( i, j = 1, \ldots, K \):
  \[
  SMJ(K) = \max_{i \neq j} \left\{ 1 - \frac{1}{\ln d(e_i, e_j)} \right\}^{-1}
  \]  
  (6)

- Modified Dunn Index (\( MDI(K) \)), adapted in [25] from the original version [55] by using the Euclidean distances; for \( i, j = 1, \ldots, K \):
  \[
  MDI(K) = \max_{i < k} \left\{ d(X_i, p) \left( \min_{i \neq j} d(e_i, e_j) \right) \right\}^{-1}
  \]  
  (7)

- Ratio of within cluster sum of squares to between cluster variation (\( WCBCR(K) \)) [20]; for \( i, j = 1, \ldots, K \):
  \[
  WCBCR(K) = \sum_{i=1}^{K} \sum_{k=1}^{K} d^2(e_{ik}, x_i) \left( \sum_{i=1}^{K} d^2(e_{ik}, e_{ij}) \right)^{-1}
  \]  
  (8)

- Intra-cluster index (\( IAI(K) \)) [36], related to the basic distances:
  \[
  IAI(K) = \sum_{i=1}^{K} \sum_{k=1}^{K} d^2(e_{ik}, x_i)
  \]  
  (9)

- Inter-cluster index (\( IEI(K) \)) [36], related to the distances to the pooled scatter \( p \):
  \[
  IEI(K) = \sum_{k=1}^{K} n_i d^2(e_i, p)
  \]  
  (10)

Each clustering indicator can be applied to a data set formed by using either time domain data or data defined in other vector spaces. As discussed in section 4.1, representing the clustered groups with their time-domain patterns enables direct comparison among the clustering outcomes. Table 2 summarizes the use of the clustering validity indicators in various literature references.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>( MIA(K) )</td>
<td>[2][10][12][13][20][21][22][25][32][37][41][42][56]</td>
</tr>
<tr>
<td>( CDI(K) )</td>
<td>[2][12][13][20][21][22][25][32][37][41][42][56]</td>
</tr>
<tr>
<td>( VRC(K) )</td>
<td>[12][13][19][20][22][25][32][39][43]</td>
</tr>
<tr>
<td>( IAI )</td>
<td>[19][36][43]</td>
</tr>
<tr>
<td>( IEI )</td>
<td>[36]</td>
</tr>
<tr>
<td>( MDI )</td>
<td>[25]</td>
</tr>
<tr>
<td>( SI )</td>
<td>[12][13][22][25][32][41][42]</td>
</tr>
<tr>
<td>( SMJ )</td>
<td>[12][13][20]</td>
</tr>
<tr>
<td>( WCBCR )</td>
<td>[20]</td>
</tr>
</tbody>
</table>

5. POST-CLUSTERING PHASE

The clustering results are used to set up the customer class representative load patterns, called load profiles. Generally the load profiles are expressed in absolute terms, that is, the vertical axis is expressed in power units. This is done to make the interpretation of the load profile simpler to the reader. Alternatively, the load profiles could remain in relative terms, but the reference powers associated with each of them should be clearly defined.

The load profiles are not necessarily given by the centroids resulting from the application of the clustering algorithm. In particular, if the clustering is done by using features different from the time domain data, the relevant outcome of the clustering process is the group formation, as indicated in section 3.2, but the final load profiles are in any case determined by using the load patterns expressed in the time domain. The same recalculation occurs if the centroids are calculated by simple average of the RLPs (without considering their reference power).

Another important aspect for load profile formation is that the class representative load patterns that can be built on the basis of the clustering results could be referred to the only customers subject to on-site measurement, that may correspond to a limited number with respect to the entire customer set. In this sense, more refinements would be necessary to build the load profiles representing the whole population of customers. The final load profiles can be obtained by properly rescaling the class representative load patterns, taking into account not only the reference power, but also other scalar factors introduced for the purpose of reproducing with the load profiles the overall energy consumption of the entire customer set. This determination requires...
availability of further measurements at the substation level, knowing the exact location of all the consumers served by that substation, in order to match the actual consumption pattern by using the load profiles in the best way possible. The evaluations can be carried out by exploiting data fitting techniques.

The final load profiles are used by suppliers and authorities to formulate and check the effects of dedicated tariff offers for the actual consumers. In addition, the load profiles provide the basis for making further assessment, for instance to test the revenues that could come to the supplier by modifying the tariff offer, or to estimate the energy not served after an interruption affecting a known portion of the network.

For tariff purposes, the load profiling system has to be readily adaptable to incorporate the effects of changing the number of consumers and their characteristics (for instance, for a contract power variation). In particular, clear solutions have to be set up to incorporate in the load profiling system the presence of new consumers [2]. When new consumers are added, their attribution to one of the existing consumer classes can be done on the basis of its estimated load pattern based on initial estimation of its type of application (e.g., based on external features). Then, the attribution of the new consumer to the existing classes can be refined according to measurements to be carried out in the first period of the consumer connection to the supply system, by determining with respect to which centroid the new load pattern has the lowest distance. It can be also noted that, after including new consumers, the load profiles should be periodically updated (e.g., once a year) in such a way that the time integral of the overall load pattern curve referred to a customer class in the various loading conditions matches the total energy consumed by all customers belonging to that customer class. The attribution of new consumers to the existing customer classes can be assisted by the use of classification algorithms such as the one used in [10], based on C5.0 [57]. Classification of the business activities into their most probable clusters has been carried out in [19] by using Probabilistic Neural Networks (PNN) [58].

6. SUMMARY OF THE CLUSTERING RESULTS

The clustering methods tested by various authors provide useful information on load pattern clustering. In some papers the methods were compared on the basis of the clustering validity indicators. The uniform definition of the clustering validity indicators, according to which lower values correspond to higher validity, makes it possible to observe and rank the methods by testing the results obtained for various numbers of final clusters. The dependence of the indicators on the final number of clusters implies that the methods can be compared only by considering the same number of clusters formed. A ranking of the methods can however be carried out by checking the robustness of the results, namely, the fact that the same method provide the lowest values of the clustering validity indicator for different indicators across different final numbers of clusters.

In some cases, the best number of clusters formed is determined by tracking the evolution of specific indices (such as the entropy content [32]), and identifying the presence of maximum conditions for these indices. However, for large data sets the best number of clusters obtained in this way could be excessively high with respect to the needs of the electricity suppliers, according to which the final number of customer classes has to be relatively low.

The convenience of adopting a clustering method rather than another one also depends on the nature of the data set. From extended comparisons made on electrical load pattern data sets containing some hundreds of patterns, it emerged that some methods regularly provided more effective results in terms of clustering validity. In particular, among the most classical methods, the hierarchical clustering method with average linkage criterion has been the one showing remarkably good performance [12][20][21][25]. Comparable performance with respect to HC has been shown by the version of the FDL method introduced in [2], by the IRC method [41] merging the complementary characteristics of HC and FDL, as well as by SVC [42] (especially for low numbers of clusters) and by the Renyi entropy-based method specifically developed in [32] to deal with the load pattern classification problem.

Generally, the most interesting results have been shown by exploiting methods exhibiting significant ability to isolate the outliers appearing in the data set. Other methods like k-means exhibits some trend to create more uniform groups, with lower attitude to single out uncommon load patterns. Finally, interesting perspectives have been opened by the use of non-Euclidean metrics, such as the ones adopted in the IRC method and to create the variants of the Renyi entropy-based method. A challenging aspect is the assessment of the potential of non-Euclidean metrics to be used in the clustering procedures to deal with the specific problem of load pattern classification.

7. CONCLUSIONS

Research on electricity load pattern classification has shown that classical clustering methods such as k-means and some variants of the hierarchical clustering are not exhibiting the best performance in forming the customer groups by singling out the outliers existing in the data set. Among the clustering algorithms tested, the FDL, SVC and Renyi entropy-based methods emerged as promising options, with results comparable with the hierarchical clustering executed with the average linkage criterion.

The adoption of a clustering method able to isolate the outliers opens the question on how to handle the outliers, especially when their number is non-negligible with respect to the total number of clusters formed. In this respect, the supplier (or the authority) is the decision maker establishing the treatment of the outliers, for
instance forming specific load profiles of individual customer classes, or including them in other customer groups.

One of the directions to be explored for enhancing electrical load pattern classification refers to the formulation and testing of suitable techniques for handling very large amounts of data gathered from many consumers, as the data made available by the extended adoption of smart metering technologies. Further directions include the incorporation of the effect of possible demand response actions [38] or the application of real-time pricing concepts [59] in the customer class formation, as well as further exploitation of non-linear metrics within the clustering algorithms to make the clustering process more efficient.

8. REFERENCES


