Unveil the Shape: Data Analytics for Extracting Knowledge from Smart Meters

Gianfranco Chicco, *Politecnico di Torino, Italy*, Diego Labate, *e-distribuzione, Roma, Italy*, Antonio Notaristefano, *Politecnico di Torino, Italy*, Federico Piglione, *Politecnico di Torino, Italy*

Received: 18 December 2019; Accepted: 20 January 2020; Published: 18 April 2020

***Abstract*—The increasing availability of data gathered from smart meters makes it possible to develop specific data analytics procedures to extract knowledge from data. The analysis of huge amounts of data coming from millions of users needs appropriate platforms and efficient algorithms. The platform SHAPE (Statistical Hybrid Analysis for load ProfilE), set up by e-distribuzione, has enabled to determine, for the first time in Italy, the consumption classes at national level based on the behaviour of low-voltage customers. This paper discusses the data handling process in SHAPE, based on customer sampling by macro-categories, definition of consistent time periods, and clustering-based load profiling, and addresses specific applications such as customer classification, load prediction, probabilistic aggregation of residential loads, and sharing of the energy not supplied by macro-categories of users.**

***Index Terms*—load pattern shape, load profiling, data analytics, smart metering, customer categories, residential consumers, industrial users.**

I. Introduction

THE progressive introduction of smart meters around the world is making available several data gathered from end-use customers at resolutions of 60 minutes or lower. This leads to a growing interest in the development of data interfaces and applications that extract information from the time series containing load curve data. Managing the continuous flow of data from millions of customers is quite challenging. Appropriate tools are needed to increase the effectiveness of data management carried out by the subjects involved in the electricity distribution and supply chain, with the aim of providing better customer services. A recent report of the European Smart Grids Task Force [1] contains an explicit recommendation to the European Union to create a smart meter roadmap, in which smart meters (also with high-resolution time intervals) should be able to meet the requirements of future markets, providing a modular and flexible architecture for the metering infrastructure. In this context, the smart meter outcomes contain useful information on the shape of the electricity consumption. This information is remarkably important to partition the customers into a number of classes that represent how electricity is used during time from different customer groups.

The above concepts have been considered to create a dedicated platform called SHAPE (Statistical Hybrid Analysis for load ProfilE), in which data analytics has been used to extract knowledge from the data gathered by the smart meters installed by e-distribuzione in Italy. The outcomes of the platform SHAPE have revealed, for the first time in Italy, the consumption classes of the low-voltage customers at the National level based on the consumption time series. The entire load-related dataset was processed automatically and collected in anonymous form. None of this information was correlated to the natural person/customer and did not allow identification in any way.

The main characteristics of SHAPE have been presented in [2], which describes the architecture and data warehouse, summarises the contents of the main modules (hourly energy analysis, customer classification, load prediction, and non-technical losses detection), and provides indications on the first two modules. In particular, the load geographical aggregator and the load pattern viewer bring extensive visualisation capabilities to the operators. Other significant subdivisions, such as the position on the Italian territory (e.g., north, centre, south), the location of customers (e.g., urban, suburban or rural) and the temporal distribution (e.g., day in the week and seasonality) may be identified and detailed. The load prediction and non-technical losses detection modules have been described in [3]. These modules make short-term and medium-term predictions available for any customer aggregation, and support the operators in the identification of possible fraudulent customers from the assessment of their metered data and the comparison with specific attributes that denote the possible occurrence of frauds.

The contents of this paper complement the previously presented information by discussing the scientific aspects referring to a number of contents implemented in the SHAPE platform. The results obtained make it possible to develop a service dedicated to the definition of consumption classes depending on the shape of the load curves, rather than merely based on commercial parameters (which do not generally represent groups of users having in common the same actual shape of the electricity consumption [4]). The emphasis on the shape of the load curves is the dominant aspect that led to the implementation of the SHAPE portal. Consumption classes are determined by constructing load curves representative of individual users at various times of the year and in various days of the week. Furthermore, important information for the management of the electrical system comes from the study of the aggregation of load curves for the determination of load profiles for each consumption class. The load profiles can be defined in a deterministic or probabilistic way, in the latter case also representing the uncertainty associated with the definition of the mean values. In this context, it is of particular interest to evaluate the probabilistic distributions of the aggregate residential customer load for each quarter of an hour. The results obtained also allow determine the aggregate energy demand for the different classes of users in appropriate time slots, for interacting with the electricity markets.

Compared to the assessments concerning the study of the operating conditions of the electricity grid, the use of load profiles is useful for reconstructing the aggregate load curves in areas where the load curves are not available for all the customers. Load profiles can be used in the analysis of electrical distribution networks, for the estimation of power flows and losses. The calculation of the correlations between load profiles and the load curves of customers belonging to different commercial categories can be used to quantify the impact of the various commercial categories in the definition of load profiles. From the point of view of reliability calculations, the use of load profiles is also appropriate for estimating the energy not supplied following interruptions of a defined duration that occur in specified time periods.

The next sections of this paper are organised as follows. Section II deals with the organisation of the data analysis in SHAPE, with the partitioning of the customers into macro-categories, the selection of a statistically significant number of samples, and the definition of consistent time periods in which the consumption characteristics of the customers belonging to the same macro-category are relatively similar. Section III describes the process of customer classification, in which clustering algorithms are run for each macro-category to identify groups of users with similar shapes of the load curves, then a decision tree-based classifier is used to form the final customer classes. Section IV presents the integrated framework created to carry out medium-term load prediction, which is used also for the generation of parts of the load curves that replace absent or unreliable metered data. Section V discusses further applications based on the metered data, to perform probabilistic analysis of an aggregation of residential loads, and to determine a conventional partitioning among the macro-categories of the energy not supplied during an interruption period. The last section contains the concluding remarks.

II. Preliminary Analysis: Setting up the Framework

The study of the load curves associated with low-voltage (LV) customers starts with the evaluation of the size of the sample of customers to be analysed through statistical sampling procedures. This assessment is carried out within appropriately identified macro-categories, which concern both consumption and local producers. These macro-categories include households (residents); households (non-residents); general building services; transport; agriculture; industry; commercial; public lighting; heat pumps; and generation (producers/prosumers).

The macro-categories are defined (and the minimum size of the sample of users to be analysed is evaluated) to consider them separately in the application of clustering procedures aimed at partitioning the load curves[1] of LV customers. The data used have been preliminarily subject to consistency checks, in order to verify their completeness and the absence of anomalous information (bad data). The consistency check returns a code associated with each data provided, which identifies the data as reliable or indicates the possible source of unreliability. The missing or bad data will be either ignored in the data analysis procedures for load profiling, or will be replaced with realistic data obtained through the application of data prediction techniques like the ones indicated in Section IV.B.

Once the available data have been defined, two specific aspects are considered, in order to select meaningful subsets of data having in common some general properties of the load curves. The first aspect is the extraction of a statistically significant sample of customers that can represent the characteristics of the whole population of customers belonging to the same macro-category. The second aspect is the definition of consistent time periods during the year, in which customers belonging to the same macro-category exhibit similar behaviour in their electricity use. These aspects are detailed below.

A. *Sampling of the Population of Customers*

The total number of users supplied by the LV distribution network is remarkably high. Gathering the data of the temporal evolution of the electric load curves for all users is rather impractical for the purpose of load profiling. Thereby, it is possible to apply statistical analysis techniques used in different sectors to evaluate the number of users for whom systematic load curves can be predicted over time, in order to obtain results statistically significant for the entire customer population. In particular, it is possible to refer to statistical theories that allow determine the number of subjects to be kept under observation within a population of subjects subdivided into a set of macro-categories defined a priori. The objective is the identification of the number of metered customers to be considered, in order to obtain representative results of the categories of consumption of the Italian population of users in a statistical sense. In this case, it is useful to adopt the stratified sampling method formulated by Jerzy Neyman [5], in which each stratum corresponds to a macro-category. This method is applied with a multi-step procedure, structured as follows:

1. *Step 1*: definition of the macro-categories. Let us denote with H the number of macro-categories, and with N_h the number of customers belonging to each macro-category $h = 1,\ldots, H$.
2. *Step 2*: selection of the characteristic variable to be evaluated. This variable has to be available for all the customers. Two variables could be considered for statistical evaluation, namely, the contract power and the annual active energy, available from the Company databases. Conceptually, these two characteristic variables are scarcely or not at all interrelated, because in general customers with the same contract power have different utilisation[2]. For the determination of the load profiles, it would be preferable to take into account the *annual active*

[1] This paper indicates as “load curve” the time series that contains the sequence of average active power values inside a time period, for a given time step. This term is used with a general meaning and has to be replaced with “generation curve” when the time series refers to generated active power.

[2] The *utilisation* is the ratio between the energy used in a given time period (typically one year) and the contract power, and is measured in hours.

energy, since the information obtained from the load curves concern how the energy is used over time.

3. *Step 3*: data collection to obtain, for each macro-category defined in *Step 1*, the information regarding the characteristic variables for each user belonging to the macro-category in the entire user population.
4. *Step 4*: for each macro-category $h = 1,\ldots,H$, calculation of the statistical parameters (mean value μ_h and standard deviation σ_h) of the characteristic variable.
5. *Step 5*: determination of the statistically significant total number of points *n*, for a given confidence probability with associated multiplier *k* of the estimated standard deviation, and per cent amplitude *d*% of the confidence interval referring to the mean value of the characteristic variable [2] (contract power in Fig. 1a, and annual active energy in Fig. 1b):

$$n(d\%,k) = \frac{\left(\sum_{h=1}^{H} N_h \sigma_h \sqrt{\frac{N_h}{N_h-1}}\right)^2}{\sum_{h=1}^{H} \frac{N_h^2\sigma_h^2}{N_h-1} + \left(\frac{d\%}{100\,k}\sum_{h=1}^{H} N_h \mu_h\right)^2} \qquad (1)$$

For example, by using the annual energy as the characteristic variable, with 99% confidence probability ($k = 2.58$) and $d\% = 5\%$ the total number $n(d\%,k)$ is about 17,000 (Fig. 1b). This number can be considered as a minimum acceptable reference. For each macro-category *h*, the number of users n_h to be observed is then determined from the equation [6]:

$$n_h = n \frac{N_h \sigma_h \sqrt{\frac{N_h}{N_h-1}}}{\sum_{m=1}^{H} N_m \sigma_m \sqrt{\frac{N_m}{N_m-1}}} \qquad (2)$$

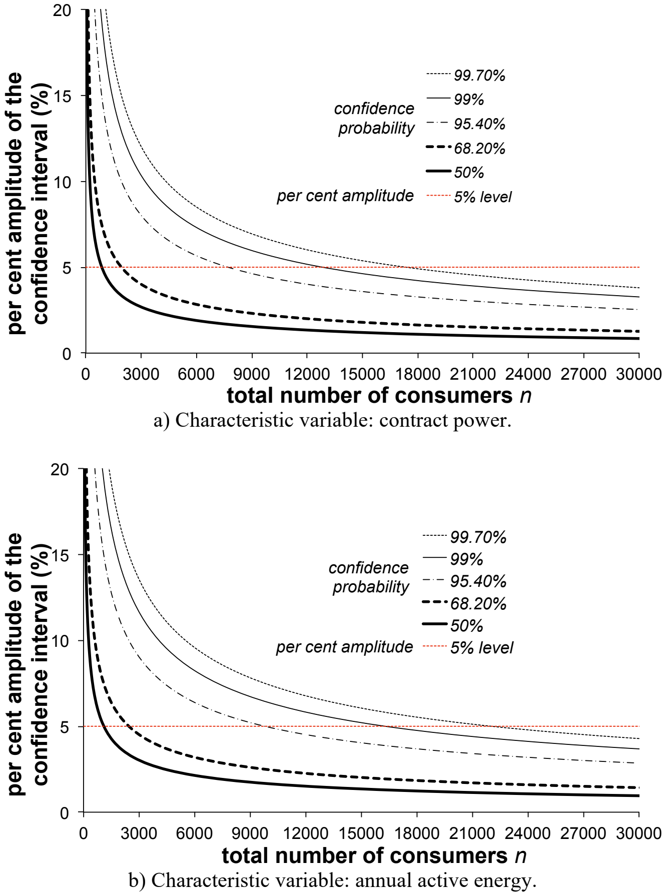
from which it is apparent that more users have to be observed for the macro-categories with many users and higher diversity in the values of the characteristic variable. In addition, each macro-category can be partitioned into sub-categories on the basis of the characteristic variable, by replicating the stratified sampling approach at the individual macro-category scale.

6. *Step 6*: estimation of the significance of the results, for each stratified sampling obtained with different values of the total number of observed users *n*, considering a given confidence probability with the corresponding multiplier *k*.

In practice, if the number of available samples is greater than the minimum values, the statistical significance of the study is higher than the one referring to the minimum values. From the foregoing considerations and from Fig. 1, it can be seen that by choosing a total number of samples equal to 20,000 an acceptable significance is obtained both in terms of contract power and in terms of annual active energy. Table I shows the partitioning into macro-categories of the minimum number of users when the total number of samples is 20,000.

Inside the macro-categories, the data available contain the mean values and the standard deviations of the annual energy for given contract power ranges. Therefore, the stratified sampling procedure gives the minimum number of samples inside each range for each macro-category. Table II shows the results for the macro-category *Industry*.

After having determined the minimum number of users for each macro-category, it is necessary to select the users to be considered as samples. Conceptually, the probabilistic distribution of samples in a macro-category should cover the probability distribution of the population of the same macro-category. Having available all the values of the considered characteristic variable (annual energy) for the entire population, it is possible to construct the cumulative curve of the values of the characteristic variable, having in the abscissa the values of the variable and the ordinates defined in the interval [0,1]. Then, the selection of the samples occurs by extracting, for each sample, a random number from a uniform distribution in [0,1], entering the value on the vertical axis, and choosing the sample from the abscissa of the point found on the CDF. This extraction is continued until the minimum number of samples is reached.



a) Characteristic variable: contract power.

b) Characteristic variable: annual active energy.

Fig. 1. Results of the stratified sampling.

TABLE I. MINIMUM NUMBER OF USERS FOR SAMPLING.

Macro-category	*Characteristic variable*	
	contract power	*annual active energy*
households (residents)	923	1954
households (non-residents)	555	1809
general building services	759	1967
heat pumps	2	2
agriculture	422	2464
commerce	12531	7885
transport	329	607

industry	878	1619
producers	186	819
public lighting	3417	874
TOTAL	*20000*	*20000*

TABLE II. PARTITIONING OF THE MINIMUM NUMBER OF SAMPLES FOR THE USERS OF THE MACRO-CATEGORY *INDUSTRY*.

Contract power range	*Number of users*	*Annual energy mean [kWh]*	*Annual energy standard deviation [kWh]*	*Minimum number of samples*
0 - 6 kW	178,113	2,258	3,696	112
6 - 15 kW	99,150	8,906	9,363	158
15 - 30 kW	51,072	21,779	20,770	180
30 - 40 kW	17,583	38,777	45,917	137
40 - 55 kW	18,808	50,767	41,525	132
> 55 kW	41,339	129,978	128,236	900
TOTAL	406,065	23,167	58,102	1,619

B. Creation of the Consistent Time Periods

The traditional partitioning of the time periods for representing the electrical behaviour of the customers is based on the seasons of the year. Then, further distinctions are typically introduced by taking into account, for each season, weekdays and weekend days, in the latter case with possible partitioning into Saturdays and Sundays. Finally, some special days [7], identified a priori from the calendar or discovered as days with anomalous load patterns [8][9] can be extracted out of the weekdays and are in general removed or associated with Sundays. However, this intuitive partitioning of the time periods does not take into account the fact that some distinctions (e.g., Saturday and Sunday) may be effectively relevant for some macro-categories and not for other ones. In addition, the effect on the shape of the load pattern in different periods of the year may change for different types of customers.

From the load pattern data, a data-driven extension of the concept of partitioning the time periods relevant for load pattern representation leads to the definition of what are identified here as *Consistent Time Periods* (CTPs). The CTPs are groups of days of the year in which the customers of the same macro-category exhibit a similar behaviour in their electrical load patterns. For each macro-category (or class, if there is a further partitioning into customer classes) the load patterns belonging to the same CTP can be represented by using a single average pattern (called load profile). In this way, the entire set of customers is described through a reduced set of load profiles. Moreover, the CTP identification assists the operators in the formulation of similar tariff structures or market offers for each CTP [4].

The variability of the load pattern shapes for the individual customers in a macro-category is generally high. However, the final number of CTPs has to be relatively limited, as the main trends from each macro-category have to be captured, skipping some details that could be valid only for a minority of customers inside the macro-category. In fact, these details will be explained by using specific clustering procedures to create the customer groups for each macro-category. For this purpose, the initial assumption is to consider a partitioning with A alternative time periods. For example, considering an industrial customer whose annual load pattern is shown in Fig. 2, the A = 36 alternative time periods represented in Fig. 3 are given by the combinations of the 12 months with three types of day for each month (i.e., weekdays, Saturdays and Sundays). This choice avoids that the automatic procedure groups together combinations of a few individual weekdays belonging to different months, because such a grouping would be poorly meaningful and would lead to excessive diversification of the grouping possibilities for the time periods. Conversely, the possibility of grouping the weekdays belonging to different months is practically relevant and is allowed by the initial time interval partitioning chosen. Without loss of generality, for testing the procedure to form the CTPs, the weekdays belonging to the same month are considered all together. Specific refinements could be done by taking a less compact definition of the alternative partitioning of the time periods (e.g., with weekly or two week-based initial time periods), without changing the conceptual framework presented here.

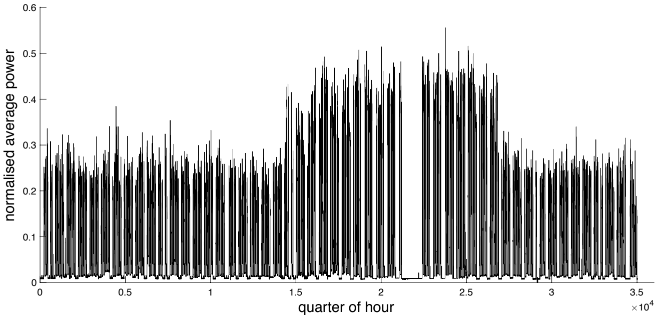


Fig. 2. Annual load pattern with data at 15-min time step for an industrial customer.

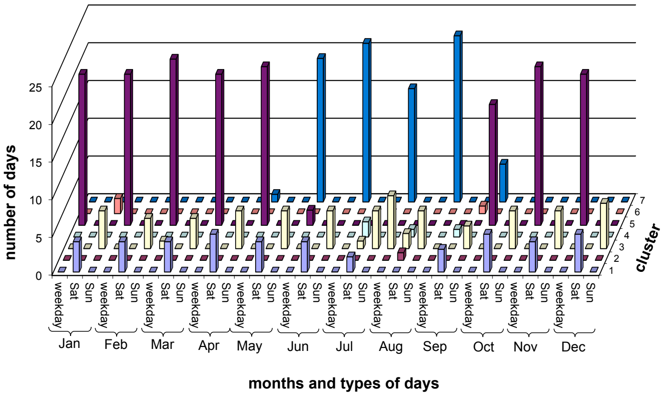


Fig. 3. Alternative time periods for an industrial customer.

As a particular case, specific calendar days representing intra-week holidays for the year under analysis are identified and associated to the Sundays of the respective month. For dedicated portions of customers included in a territorial jurisdiction in which local holidays can be identified, these local holidays may be taken into account as well, even though their effect on the macro-category for larger territorial groupings is likely to be very limited.

An automatic procedure to form the CTPs has been developed for SHAPE. A greedy algorithm has been used, as an optimisation procedure based on integer programming could be intractable in large-scale applications with a number of CTPs unknown a priori [10]. The procedure is composed of two stages. The overall scheme is shown in Fig. 4. In the first stage, the individual load patterns of each macro-category are subject to a dedicated analysis, which includes load pattern clustering and the formation of a binary vector that contains the information on the most significant alternative time periods. In the second stage, the binary vector representations of all customers belonging to the same macro-category are processed together to determine the final grouping of the time periods. Fig. 5 shows an example of CTPs for one year, found for the macro-category *Industry*.

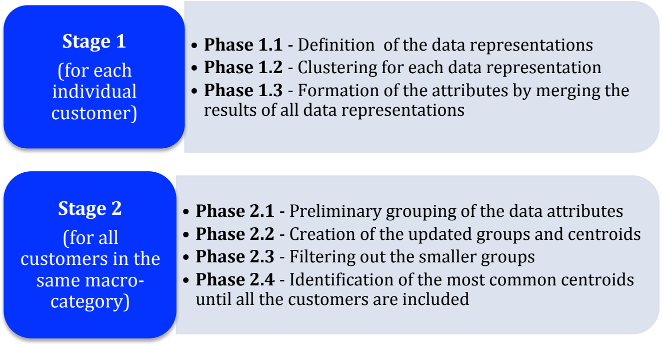

Stage 1
(for each individual customer)
• Phase 1.1 - Definition of the data representations
• Phase 1.2 - Clustering for each data representation
• Phase 1.3 - Formation of the attributes by merging the results of all data representations
Stage 2
(for all customers in the same macro-category)
• Phase 2.1 - Preliminary grouping of the data attributes
• Phase 2.2 - Creation of the updated groups and centroids
• Phase 2.3 - Filtering out the smaller groups
• Phase 2.4 - Identification of the most common centroids until all the customers are included

Fig. 4. Two-stage procedure for the definition of the Consistent Time Periods.

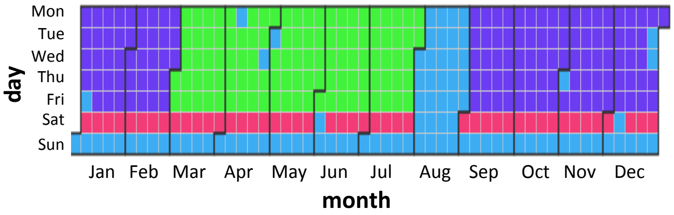


Fig. 5. Example of Consistent Time Periods for the macro-category *Industry*.

III. Shape-Based Customer Classification

The classification process implemented in SHAPE includes the partitioning of the overall set of customers for each macro-category into a number of classes, carried out on the basis of clustering results and the application of a classification model driven by supervised training. Then, new customers with unknown class may be presented to the model to be associated to one of the existing classes.

The objective of this section is to indicate some algorithms with adequate performance used into the SHAPE platform on the basis of approximately 100,000 customers, using for each customer a typical daily load curve with 15-min data.

Within each CTP, each customer participates in the partitioning process with a single typical load curve (TLC), obtained by calculating the average of the data found for the corresponding days for each quarter of an hour.

The entire load curve categorisation process can be summarized as follows [11]:

A) *Data acquisition and check*: includes the identification and correction (or elimination) of the unacceptable (or bad) data points.

B) *Pre-clustering phase*: the load curves are represented through a specific number H and type of characteristic quantities, then the input matrix for the clustering procedure is built, formed by M×H components (with the customers on the m = 1, ..., M lines, and the entries of the characteristic variables on the h = 1, ..., H columns).

C) *Clustering execution*: the clustering procedure chosen is executed, assigning each load curve to a group (cluster); using cluster composition and 15-min input data, the cluster centroids are formed; the comparison between the results of the different clustering algorithms can possibly be carried out by using clustering validity indicators [11];

D) *Post-clustering phase*: includes the activities related to forming the consumption (or local production) classes and determining the corresponding load profiles.

A. Load Curve Clustering

For the applications developed in SHAPE, the clustering algorithms used have to deal with huge amounts of data. This impacts both on the computation time and with the number of data to be stored in the memory of the computer.

A.1. Definition of the data used

With the same data describing the quarter-hour trend of local loads or generations, it is possible to define different sets of characteristic quantities to be supplied as inputs to the clustering procedure. The definition of these sets are based on different concepts:

a) Definition of data in the *time domain*: in the simplest case, the original metered data can be used directly. Alternatively, data pre-adaptation can be performed over time (with uniform distribution, or uneven distribution [12]), or over the amplitude (original amplitudes; normalization with respect to the maximum value of the typical curve, so that all values fall within the interval [0,1]; scale factor equal to the average value of the load curve; scale factor equal to a reference parameter, for example the contract power or the peak value of the curve in the period considered; min-max normalisation). In SHAPE, the selected way has been the normalisation with respect to the contract power, to highlight the importance of the shape of the load curve shape, together with the average level of consumption.
b) Definition of data in the domain of *other variables*: in this case, the goal is generally to obtain a reduction in the number of data to be stored for each user, so as to also reduce the calculation times of the clustering algorithms. Among the various techniques that can be used, it is possible to indicate techniques based on the transformation of data in a specific domain, including projection methods, through which the set of input quantities is projected onto a suitable multi-dimensional space [13,14], methods elaborated in the frequency domain, based on the coefficients of the Fourier series [15] or from their processing [16], or using wavelets [17]; methods based on representing the information with the use of shape factors [4,18,19] calculated by subdividing for example the day in the night and day periods with determination of interval factors [0,1], or on the use of "alphabetic" components obtained from symbolic aggregate approximation (SAX) method [12,20].

An important aspect to compare the solutions is the *common point* used for the analysis: regardless of the type of data used, the output of the clustering algorithm consists of a vector that contains, for each input curve, the number of the group to which the curve has been assigned by the clustering procedure.

The results obtained from the various clustering techniques depend in particular on the notion of *distance* used to evaluate how much the load curves are different with each other. The use of distances leads to a purely quantitative technical evaluation, which does not take into consideration qualitative or external considerations on the affinity of a load curve with respect to another. The calculation of the distance can refer to different types of distances. Among them, the generalised Minkowski distances consider the H values of load curve like H components of a vector. Taking two vectors $\mathbf{x}^{(i)} = \left\{x_h^{(i)}, h = 1, \dots, H\right\}$ and $\mathbf{x}^{(j)} = \left\{x_h^{(j)}, h = 1, \dots, H\right\}$, the Minkowski distance is defined by using the exponent p, where the classical Euclidean distance corresponds to $p = 2$:

$$d^{(p)}\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \left(\sum_{h=1}^{H} \left(x_h^{(i)} - x_h^{(j)}\right)^p\right)^{1/p} \tag{3}$$

In spite of various attempts, there is still no confirmed clear evidence that the use of the Euclidean distance can be less effective than another distance for the problem under analysis. The Euclidean distance is then used, and is extended to formulate distances between a vector and a group, between groups, or the inter-distance within a group [4]. These distances are also used to determine the performance of clustering methods through the use of appropriate clustering validity indicators [11].

A.2. Characteristics of the clustering methods

The choice of the most suitable clustering algorithm for the analysis to be performed (and of the metric to be considered within the algorithm) depends on multiple aspects, including:

a) The possibility of setting up the number of clusters to be obtained.
b) The ability of the clustering method to isolate the curves that show uncommon trends (outliers).
c) The solution mechanism, based on a single stage that uses all the input parameters, or on several stages, each of which depends on different parameters.
d) The computation time of the solutions and its progression as the number of input curves changes.
e) The occupation of data memory, which can become a limiting factor in the execution of clustering procedures for the analysis of a very large number of curves.

A group of methods suitable for the analysis of large datasets have been considered. The main feature that distinguishes the clustering methods takes into account the procedure followed to identify the groups. The clustering algorithms applied in SHAPE are then distinguished into hierarchical and non-hierarchical methods. Furthermore, an efficient *hybrid* clustering algorithm has been implemented to merge the benefits of the two types of clustering.

A.2.1. Hierarchical methods

These methods explore the set of possible partitions by identifying nested grouping structures. The agglomerative version has been used. Initially each element constitutes a cluster of its own. The algorithm then merges the most similar clusters, based on the minimum distance calculated between the sets. The distance to be minimised depends on the so-called *linkage criterion* (e.g., nearest neighbours, farthest neighbours, centroid, average, or Ward [21]). The average linkage criterion has been found to be the most appropriate for the specific application. The algorithm continues to merge the clusters for increasing distances until the desired number of clusters has been reached. In general, hierarchical methods are computationally efficient in terms of computation speed. The critical point is the high occupation of memory due to the need to construct the entire matrix that expresses the interdistances between the curves. Hierarchical methods also do not require an established initial configuration, i.e., the procedure starts by acting directly on the elements of the data set. These techniques are also sensitive to any anomalous units that are effectively isolated and allow the procedure to be stopped directly when the desired number of clusters is reached. Most of the computation time (even more than 80 per cent) is dedicated to the calculation of the distances between the pairs of entries in the distance matrix. However, significant improvements (depending on the specific hardware used) may be introduced by parallelisation of the distance calculations.

A.2.2. Non-hierarchical methods

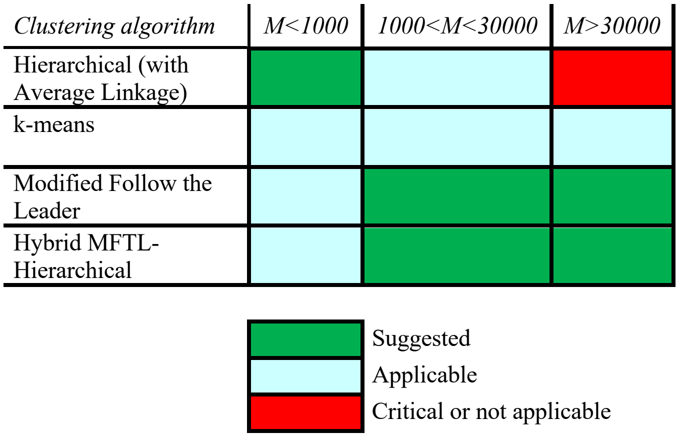
These methods are characterised by a procedure that aims to directly introduce the load curves in the groups with an initial criterion, and then optimise the partition. The algorithm is divided into two main phases: (i) a temporary partition of the load curves is identified in a certain number of clusters, generally fixed a priori, and (ii) an objective function is optimised by modifying the assignment of the elements to the groups. The identification of the optimal partition would strictly involve the evaluation of all the possible distinct assignments of the elements in the set number of groups. Since this type of operation determines a huge amount of calculations, the non-hierarchical procedures propose to solve the problem through a grouping strategy that requires the evaluation of only an acceptable number of possible alternative partitions. In practice, once the initial partition has been chosen, the units under examination are reallocated between the different groups in order to optimise the predetermined objective function. SHAPE has implemented the most widely used non-hierarchical algorithm, *k-means* [21], that requires specifying the desired number of clusters. The proposed k-means method implements the choice of initial centroids to be associated with clusters through a correlation analysis carried out on a sample of load curves belonging to the dataset. As far as calculation times and memory allocation are concerned, the implemented k-means method is suitable for the analysis of very large datasets, also containing more than 1,000,000 curves. Moreover, SHAPE has implemented the *Modified Follow the Leader* (MFTL) method [4,22], in which the execution of the first iteration involves the formation of the (undefined a priori) number of initial clusters based on the value of a threshold entered as input data. The subsequent iterations refine the cluster composition, keeping the number of clusters obtained in the first iteration unchanged. The MFTL method is of interest because of its calculation speed and reduced requirement of allocated memory. The calculation times in the successive iterations have been improved by implementing some steps with parallel computation techniques.

A.2.3. Hybrid method

The Hybrid MFTL-Hierarchical clustering method merges the positive aspects of simplicity and computational efficiency of the hierarchical and MFTL methods, while avoiding their drawbacks (the memory allocation for the hierarchical method, and the lack of definition of the number of clusters for the MFTL). The calculations are partitioned into two stages. In the first stage, a sample of poorly correlated initial load curves having different shapes is considered, to estimate a reasonable value of the threshold (taken at one half of the average distance between all the possible distances among the pairs of initial load curves). The first iteration of the MFTL method is run, choosing a number of initial clusters higher than the number of desired final clusters. In this way, the number of clusters is reduced to a value tractable by the hierarchical clustering, and the hierarchical clustering is then run to obtain the final clusters.

A.2.4. Scalability of the clustering methods

A key point is the scalability of the procedures when the number of load curves increases. Fig. 6 shows the applicability of the clustering algorithms in function of the number of load curves M contained in the dataset. As it can be seen, the Hierarchical clustering method has excellent performance and relatively fast calculation times when the number of load curves is relatively low, while the calculation times increase with M and may become impractical or even not applicable for $M > 30000$ because of data storage limits. Then, the suggested methods for large datasets are MFTL and the Hybrid MFTL-Hierarchical clustering.



Clustering algorithm	*M<1000*	*1000<M<30000*	*M>30000*
Hierarchical (with Average Linkage)	🟩	🟦	🟥
k-means	🟦	🟦	🟦
Modified Follow the Leader	🟦	🟩	🟩
Hybrid MFTL-Hierarchical	🟦	🟩	🟩

🟩	Suggested
🟦	Applicable
🟥	Critical or not applicable

Fig 6. Applicability of the clustering algorithms in function of the number of load curves M contained in the dataset.

B. *Determination of the Shape-Based Customer Classes*

The classification of the load curves is carried out with respect to the predefined macro-categories. Starting from the partitioning of the load curves obtained from the application of clustering procedures, the number of consumption classes is determined. Each class is represented by the corresponding load profile, composed of the reference power (sum of the contract powers of the customers belonging to the class) and the load curve expressed through normalised values with respect to the reference power. The use of the normalised load profile allows to reconstruct the aggregate load curves of a group of customers belonging to the same class to be reconstructed by multiplying the normalised load profile by the reference power of this group of customers and by any other factor that can be inserted to represent further details in the definition of the day (for example, if the day is considered an anomalous day based on its placement on the calendar as a holiday occurring on a day other than Saturday or Sunday).

From a conceptual point of view, the characteristics of a user are distinguished from the attributes, as follows:

- The *characteristics* are directly linked to the details of the load curves of the users, which therefore cannot be directly obtained without having the measured load curves.
- The *attributes* are directly related to technical or commercial data present in the users' commercial database, and may also be known for users for whom the measured load curves are not available.

In particular, it is possible to make a further subdivision considering:

a) Attributes defined for *existing users* (for which energy information derived from previous energy invoices is known); and,
b) Attributes defined for *new users* (for which there are no previous energy measures). The set of attributes is in this case limited to information not related to the use of energy.

The process of class formation and subsequent classification of new users depends on the characteristics and attributes available [23]. The classification tool used in SHAPE is C5.0, based on the algorithms developed by Quinlan [24]. The classifier is formed by a decision tree, in which the branches indicate the paths in the classification process, the nodes indicate the conditions referring to the numerical value of an attribute that determines the partitioning into paths, and the leaves represent the class resulting from the path followed. In the construction of the generation tree, starting from the whole training set, the algorithm identifies the attribute that divides more effectively the set of possible classes. The partitioning criterion is based on the highest difference of entropy occurring as a consequence of the partitioning into a set of classes.

In the solution procedure for determination of the classes, important information concern the behaviour of the user throughout the year. It is therefore necessary to get data referring to at least an entire year. From the point of view of the structure of the procedure, the classical decision-making methods, present in the literature, rank the characteristics based on the significance of the characteristics themselves. The algorithm proposed for the search of the representative classes of each macro-category is based on the analysis of two types of information: (i) the load curve analysis; and (ii) the monthly utilisation. The implemented procedure has a tree structure, so the data sets containing the two different types of information are analysed one at a time. The analysis of the load curves produces the main partition, which can be further divided by the subsequent analysis on the monthly utilisation curves. The search for groups and subgroups is carried out according to the similarities, using clustering programs, avoiding the imposition of parameters that constrain or orient the partition.

Each user can be described by a typical load profile for each CTP. The typical profile of a specific CTP is a daily load curve obtained as the average of all the daily load curves belonging to the same CTP. The calculation of the average allows filtering out any anomalous behaviour present in some days and therefore to obtain a "clean" load curve that well represents the behaviour of the user in the specific CTP.

The proposed procedure induces a partition of users belonging to the macro-category on two levels. In the first phase, the classes are identified based on the shape of the load profiles in CTPs. Subsequently, *subclasses* are defined, which are internal to the individual classes and discriminate the users according to the annual load trend, described on a monthly basis. The results indicate 22 main customers classes for the various types of loads represented in the macro-categories. For each customer class, the representative load curves obtained are validated with respect to the total energy of all the national customers included in the corresponding class. Taking into account the time partitioning given by the CTPs, the total energy is assessed and the representative load curves are rescaled to match the actual total energy of the national population. The validated representative load curves are the final *load profiles* of the customer classes and subclasses.

For example, for the class *Industry*, two classes (IND1 and IND2) have been identified, with four CTPs for each class (Fig. 7). For the first class, there are two subclasses (IND1_1 and IND1_2). Fig. 8 shows the load profiles for the two subclasses of IND1 and Fig. 8 for the class of IND2, together with the normalised utilisation for each month.

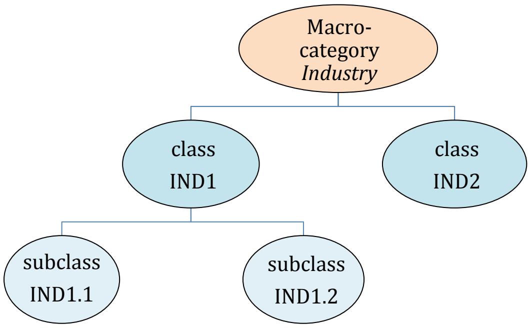
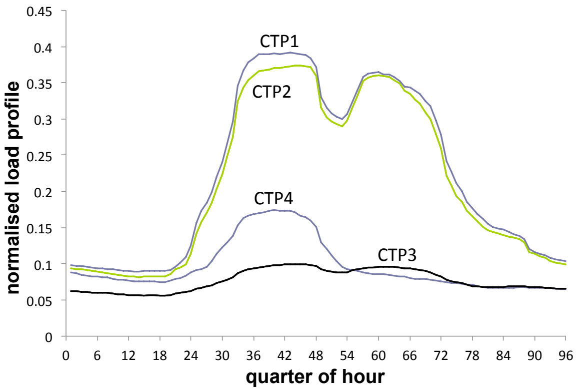
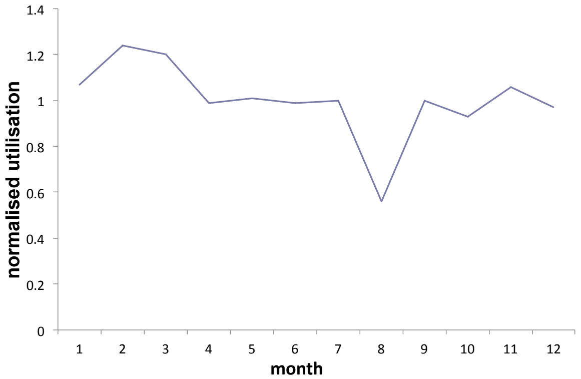


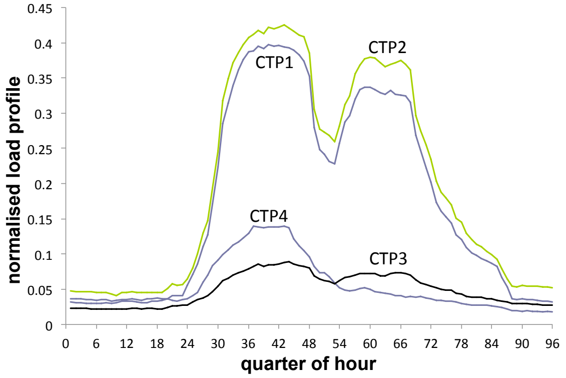
Fig. 7. Classes and subclasses for the macro-category *Industry*.



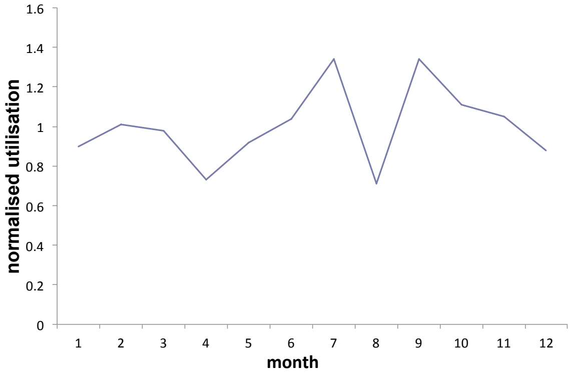
a) Normalised load profiles for Subclass IND1.1.



b) Utilisation for Subclass IND1.1.

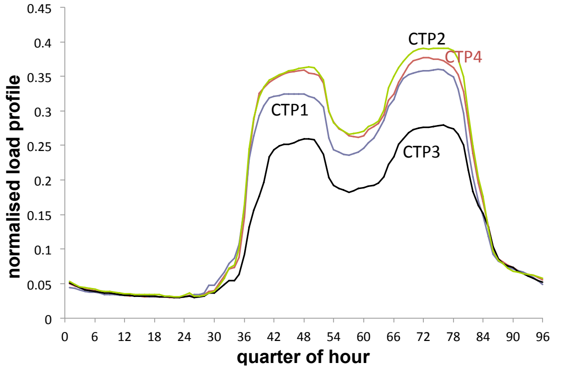


c) Normalised load profiles for Subclass IND1.2.

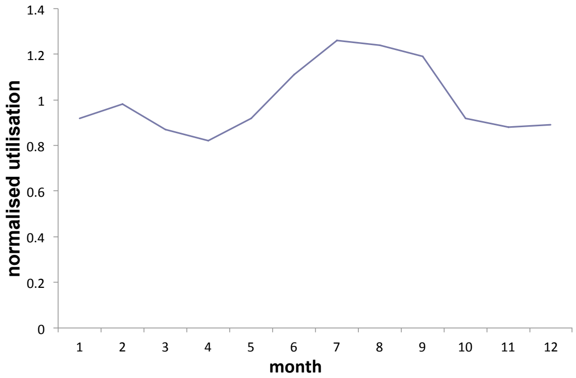


d) Utilisation for Subclass IND1.2.

Fig. 8. Normalised load profiles and utilisation for Subclass IND1.2 and Subclass IND1.2 of Class IND1.



a) Normalised load profiles for Subclass IND2.



b) Utilisation for Class IND2.

Fig. 9. Normalised load profiles and utilisation for Class IND2.

Finally, for the classification of a new user (i.e., to assign a new user to a predetermined consumption class based on a set of attributes), the energy-related attributes can be estimated by considering the type of use of the foreseen energy, or can be defined on the basis of measurements made in some time periods subsequent to the user's connection to the network.

IV. Integrated Framework for Medium-term Load Prediction and Fictitious Data Creation

An integrated approach has been established to deal with load prediction and the definition of the fictitious data for the completion of incomplete load curves with a single program. This approach is implemented in a program that performs the medium-term prediction of the load curves with the *Extreme Learning Machine* (ELM) neural network [25], with the use of the *ensemble* technique to improve the accuracy of the prediction. The ensemble technique is based on the repetition of the prediction for a predefined number of times, from which the median of the distribution of the results obtained is taken as the result. ELM has a relatively fast training [26], which makes it particularly appropriate to be used in a platform with large datasets. The same program allows the reconstruction of portions of curves with missing data, with different lengths.

Extreme Learning Machine (ELM) is a fast learning technique for Multilayer Perceptron (MLP) proposed by Guang-Bin Huang [27]. It exploits the so-called “kernel trick” to transform a linearly non-separable problem into a linearly separable one. The kernel trick is based on the Cover’s theorem [28], which states “*A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated”*.

Known since the beginning of neural networks, this method has been from time to time “rediscovered” with different names: Learning and generalization characteristics of the random vector functional-link net [29], Random Activation Weight Neural Net (RAWN) for fast non-iterative training [30], Pseudoinverse learning algorithm for feedforward neural networks (PIL) [31], and some other.

The original ELM [25] is an MLP with one hidden layer and a linear output layer. The weights and biases of the hidden layer are randomly assigned and the weights of the output layer are found by linear regression, avoiding in this way the time-consuming iterative learning. In the following years the author and other researchers continued to develop ELM in several ways (On-line sequential ELM, Evolutionary ELM, Fully complex ELM, and Multiple hidden layers ELM).

ELM was used for the electric load forecasting for its learning speed. In ensemble learning applications ELM can advantageously replace the use of parallel computing [32–34].

The primary difference between ELM and the traditional MLP is the learning algorithm. MLP is generally trained with gradient descent algorithms, such as the error back-propagation (BP) algorithm, proposed by Werbos in 1974 [35]. Despite subsequent improvements, such as adaptive learning rate and momentum, BP training is very slow. Faster training algorithms are Conjugate Gradient, Newton method, Quasi Newton and above all the popular Levenberg-Marquardt method. However, the non-iterative training of ELM is always faster than any training algorithm of MLP.

In MLP and ELM, the random initialisation of the weights causes slightly different results in different training sessions. Ensemble averaging is a machine-learning method that overcomes the problem by training in parallel several concurrent neural nets and calculating the median of their outputs. This results not only in a stable result but also in a lesser error. Two factors contribute to obtain a better result: 1) the combined effect of several MLPs compensates the different random initialisation; 2) each concurrent MLP employs a slightly different number of hidden units. Ensemble averaging with many concurrent neural nets requires parallel computing or fast training algorithms. For this reason ELM is particularly suitable for ensemble averaging.

A. Medium-Term Load Prediction

The algorithm formulated makes use of calendar variables and Consistent Time Periods (CTPs) for the definition of days in three different ways: (i) distinction between the days of the week (all 7 days, independent of each other); (ii) distinction between weekdays, Saturdays and Sundays; and (iii) distinction based on CTPs (usable only for single users or for aggregated users of the same macro-category or class for which the CTPs are defined; in fact, when users with different characteristics are aggregated, the CTPs are no longer defined in a unique way).

A.1. Implementation of the load prediction algorithm

The implementation of the load prediction requires the following steps:

1) Analysis of the load curves, to identify the autocorrelation of the temporal data, the correlation with any exogenous variables (temperature, humidity, time of day, day of the week, holidays, etc.), and the presence of trends and seasonality. The purpose of the analysis is the choice of the independent variables (regressors) to be included in the prediction model.
2) Choice of the prediction horizon (at the quarter hour, at the hour, at 24 hours) compatible with the available regressors and the formulation of the prediction model.
3) Choice of the neural network and design of its structure. The prediction model establishes the number of input and output units, while the number of other units (*hidden units*) is a design choice.
4) Creation of the data set used in the learning process (*training set*). The size of the training set depends essentially on the characteristics of the load curve studied and on the number of hidden units on the network. A training set of several weeks or months allows the learning of a high number of profiles, but could lead to generalisation issues, especially if the load does not have a well-defined periodicity. In this case a training set reduced to a shorter period of time is more effective.
5) Creation of the data set used in the verification of the forecast quality (*test set*). The data belonging to this set were obviously not included in the training set. For load prediction, the test set is generally made up of a time period immediately following the learning period.

In the application of the prediction procedure in SHAPE, the load curves are normalised to the contract power. The curves were sampled at quarter hour (96 daily samples) but are averaged to obtain hourly sampling (24 daily samples) in order to filter at least part of the noise always present in low voltage users. The main analysis tool used is the linear correlation coefficient for different intervals (lags) of delay, significant even if the actual correlation is only partially linear [36]. Fig. 10 shows the variation of the autocorrelation coefficient at different lags for the hourly load curve of an industrial customer with contract power 20 kW and annual energy 38 MWh. The highest autocorrelation occurs for lag 1, followed by lag 24, lag 2, lag 23, and lag 25. This example shows that the variability of low voltage loads is typically high, causing a steep decrease of autocorrelation at small lags and low periodicity at 24 hours. Thereby, only a few regressors with significant autocorrelation are available, and it is not easy to obtain a low prediction error.

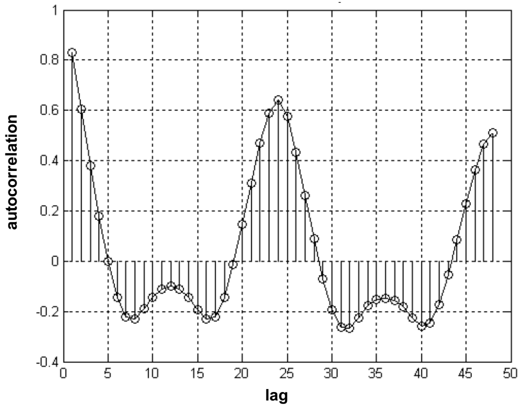


Fig. 10. Autocorrelation from lag 1 to lag 48.

A.2. Metrics to assess prediction performance

One of the most used metrics for load prediction is the *MAPE* (Mean Absolute Percentage Error). However, the major drawback of *MAPE* in the application to individual low voltage load curves is that for small values (close to zero) of the real data it can give rise to high error peaks that do not account for the soundness of the prediction. This issue does not occur for an aggregation of load curves, in which the minimum values are in general substantially higher than zero.

To mitigate this issue, it is possible to use other indicators such as *WMAPE* (Weighted Mean Absolute Percentage Error), *RMSE* (Root Mean Square Error), or *NRMSE* (Normalised Root Mean Square Error). In particular, an appropriate metric for low voltage load curves, less sensitive to low values and customized to be used with the electrical load curves, is the indicator $NRMSE_P\%$, given by the per cent *NRMSE* related to the contract power P_c of the customer. In general, if *N* real energy data are available with q data points per hour, the per cent *NRMSE* is expressed as:

$$NRMSE_P\% = 100\frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2}}{P_c/q} \tag{4}$$

where y_n is the real value of the load at data point *n*, and $\hat{y}_n$ is the predicted value.

B. Reconstruction of Continuous Load Patterns

In time series analysis, the presence of incorrect or missing data produces a lack of sequential information and therefore has a direct impact on the possibility of using automatic solutions for load prediction. In certain cases, the presence of incorrect or missing data also affects clustering procedures, especially if there is a low number of initial load curves based on which the load curves representative of typical days are generated.

The replacement of missing data may be carried out by creating fictitious data referring to time series analysis. The fictitious data must have properties similar to the time series data they represent (for example, with reference to the standard deviation of the time series).

The fictitious data are often created in order to have statistical properties similar to those of the data to which they refer. However, their temporal sequence may not entirely correspond to the time series under analysis. In fact, the calculation of statistical parameters such as mean value and standard deviation does not give information on the temporal sequence of the points that form the reconstructed curve. In the application considered in SHAPE, a fictitious data creation technique is privileged, depending on the load curve prediction.

Taking a load curve in which there is a segment of incorrect or missing data, let us assume that there are known points of the load curve before and after the segment in question. The criterion used to create fictitious data is based on the generation of a certain number of load curves starting from the last known data, using a prediction procedure. Each of the generated curves will have its own evolution. Considering the subsequent segment of known data, in the ideal case the reconstructed dummy load curve should reproduce the same trend as the known data. Since in reality this is not possible, above all because of the non-stationary nature of the data that compose the load curves, it is necessary to establish which of the generated curves connects in the best way with the known data after the segment replaced. To this end, the goodness of the connection is indicated by establishing an appropriate metric. An example of metric considers, for each generated load curve, the distance between the points generated and the known points for the corresponding times, using for the evaluation a number defined by the operator of successive points. Moreover, in order to make the initial point of connection more important than the following points, decreasing weighting coefficients are inserted into the metric. Among the generated load curves, the curve having the lower value of the indicator built according to the metric now illustrated is chosen. The prediction used to replace a portion of the curve provides results that are in any case better than using the forecast by extrapolation, since there is a link with the subsequent data. The goodness of the reconstruction of the missing data sequence depends essentially on what the previous available data allow to identify the tendency of the load curve in the missing period.

In the procedure of reconstruction of segments of particularly long load curves (for example, some days) it is necessary to insert information taken from the knowledge of the types of day that are reconstructed (weekdays, holidays or anomalous days). The "day type" attribute is considered as a calendar variable.

Let us consider a time series *f*(*t*) that contains the data gathered at the times from $t = 0$ to $t = t_0$, in which there are incorrect or missing data in the segment from $t = t_0+1$ to $t = t_1$, while the successive data for $t > t_1$ are again known [37]. Starting from the known points for $t <= t_0$, for $t > t_0$ a group of *K* time series are generated (through a method of load prediction), each of which follows the trend of the load curve to be reconstructed, extended not only to the points from t_0+1 to t_1, but also to points subsequent to t_1.

A.1. Weighting coefficients for the choice of the fittest fictitious data

The comparison between the *K* load curves generated in the prediction phase can be performed by looking for the load curve that corresponds to the best connection with the known data for $t > t_1$. The metric used to carry out the qualitative assessment of the goodness of the fitting gives greater importance to the points found in the initial part of the fitting. In particular, a system of decreasing weighting coefficients associated to the points starting from the beginning of the connection is introduced, valid for $t > t_1$ and associated with a time constant τ that represents the decay of the weighting coefficient with the growth of time beyond the point of fitting, according to the following expression:

$$w_t = e^{-\frac{t-t_1-1}{\tau}} \tag{5}$$

The operator defines the time constant τ. Indicative values, defined on the basis of the length of the connection interval (t_{max}-t_1) for which the difference between the generated curves and the actual curve is to be calculated, can be of the order of $\tau = c\cdot(t_{max}\text{-}t_1\text{-}1)$, where the coefficient *c* can for example take values between 0.2 and 0.3. The example shown in Fig. 11 indicates the evolution of the weighted coefficients calculated assuming that there are missing data up to the time step t_1 = 159, the connection starts at the time step t_1+1 = 160, and the upper limit for the evaluation of the differences between the actual curve and the generated curves is t_{max} = 180. With these values, considering the coefficient c = 0.3 the time constant results in $\tau = 0.3 \cdot (180\text{-}159\text{-}1) = 6$.

A.2. Fictitious data alignment indicator

Considering one of the *K* curves generated, indicated as $g_k(t)$, for *k* = 1, ..., *K*, the metric used to evaluate the goodness of the connection is expressed by the following indicator of *fictitious data alignment*:

$$a_k = \frac{\sum_{t=t_1+1}^{t_{\max}}\left(\frac{g_k(t)-f(t)}{f(t)}w_t\right)}{\sum_{t=t_1+1}^{t_{\max}} w_t} \tag{6}$$

Among the *K* curves generated, the choice will fall on the curve having the smaller value of the fictitious data alignment indicator:

$$a_{k^*} = \min_{k\,=\,1,\dots,K}\{a_k\} \qquad (7)$$

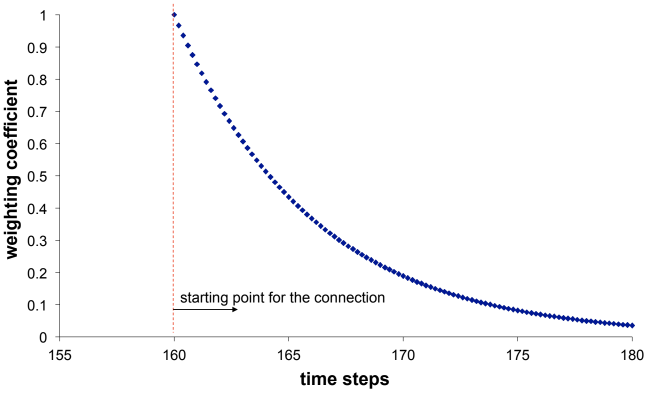


Fig. 11. Evolution of the weighting coefficients.

A.3. Examples of application to the load curve of a single user

Let us consider a customer in whose load curve there are 48 successive missing 15-min values, for a total period of 12 hours [2]. The real data of the week preceding the period of lack of values is used as the training set. The forecast is made with ELM, with H = 50 hidden units and 7 regressors {1,95,96,97,191,192,193}. The reconstructed portion of the load pattern is created by comparing K = 20 fictitious load patterns generated by the prediction procedure. Following the missing values, Z = 12 points are used to calculate the fictitious data alignment indicator, choosing in an automatic way the portion of the load curve (thick line in Fig. 12) whose data better fit the connection with the first points of the real load pattern after the time interval to be reconstructed.

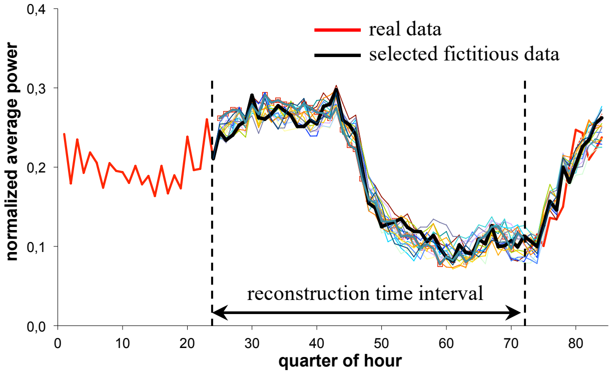


Fig. 12. Generation of 20 fictitious load curves and selection of the fittest one.

Let us now consider an aggregation of customers of mixed type, in which there are two time periods with missing data. Fig. 13 shows the results of the reconstruction, in which it may be noticed the positive effects of having separated the load curves of the weekend with respect to the weekdays.

V. Further Applications

A. Probabilistic analysis of aggregate residential loads

The assessment of the load curves for different levels of load aggregation (i.e., numbers of customers [38]) enables the operator to carry out systematic assessments of the behaviour of aggregated groups of customers. A specific point is the identification of the probabilistic distributions of aggregate residential loads as the number of customers varies. This is particularly important for residential customers [39], for which it is quite challenging to predict the shape of the load curve for individual customers, while it becomes much easier to predict the aggregated load curves at different levels of aggregation [40].

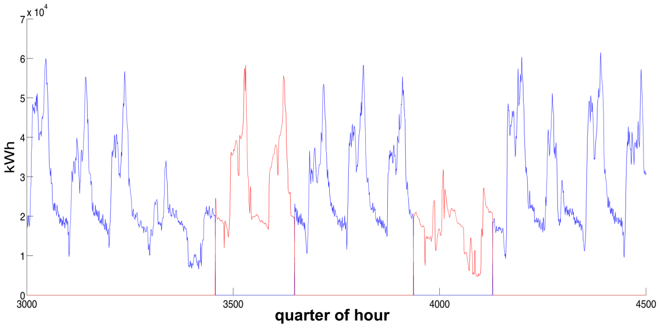


Fig. 13. Reconstruction of two time periods with missing data, considering the distinction between CTPs. Real data in blue, predicted values in red.

Let us consider the aggregation of j = 1, …, J customers. The relevant quantity observed in the study for a single customer is the energy $\chi_{j,t}$ metered at each 15-min time step t. The specific quantities (i.e., the specific mean energy $\bar{\chi}_{J,t}$, and the specific standard deviation $\sigma_{J,t}$) are determined from the statistics of the energy values, divided by the number of customers J that form the aggregation. Moreover, the standard deviation $s_{J,t} = \sigma_{J,t}/\bar{\chi}_{J,t}$ of the specific energy referring to the specific mean energy is calculated. Further quantities are obtained by considering the variations $\Delta\chi_{j,t} = \chi_{j,t} - \chi_{j,t-1}$ between two consecutive energy values.

In particular, starting from many load curves of residential customers in similar conditions (e.g., customers in urban or rural areas), by constructing aggregations with different numbers of customers it is possible to provide quantitative assessments of the behaviour of aggregate residential loads. An example is shown for the entire set of load curves available for household residents in a given territorial area (2290 load curves). Fig. 14 shows the probability density function (PDF) of the energy at each quarter of hour, drawn with classes of amplitude 0.01 kWh/15min. Since the PDF of each quarter of hour has unity area, higher peaks of the PDFs correspond to the quarters of hour in which the variability of the 15-min energy is more limited.

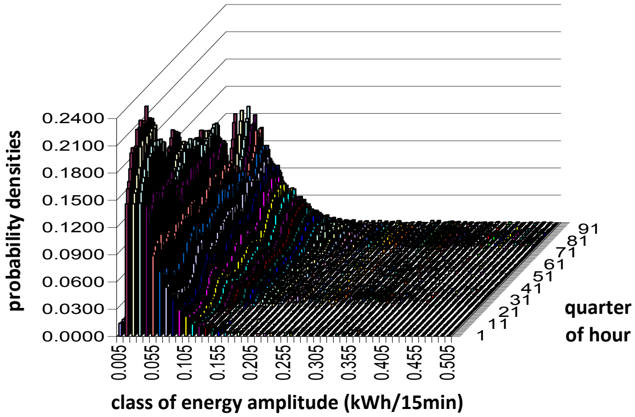


Fig. 14. PDFs of the 15-min energy for the set of residential users considered.

In the same way, it is possible to provide appropriate responses to the following questions:

1. *What are the significant parameters to represent the aggregate load curves?* Based on the results, the most suitable quantities to represent the aggregate load curves are the specific mean energy (Fig. 15) and the ratio between the standard deviation and the average value of the specific energy (Fig. 16).
2. *How the mean value and the uncertainty of the load curves depend on the number of aggregated users?* In the aggregate load curves, the average relative energy value for a given quarter of an hour has a relatively low variation for groups of users with equal characteristics (e.g., residential users in households). The standard deviation (related to the mean value) decreases instead as the number of aggregate users increases (Fig. 16).
3. *How does uncertainty in the load curves change during the day?* For residential users, uncertainty has a time dependent variation. For the example shown in Fig. 17, 101 classes of amplitude variations have been reported in Fig. 17 (from S1 to S101), each of which has amplitude of 0.006 kWh/15min, with 50 classes for positive variations, a class for variation around zero, and the other 50 classes for negative variations. The first class (S1) also accumulates the residual negative variations with amplitude greater than 0.3 kWh/15min. Moreover, a significant quantity is the standard deviation of the 15-min energy *variations* between successive quarters of an hour (Fig. 18), which for the same example is strongly correlated (i.e., correlation coefficient 0.904) with the specific mean energy.

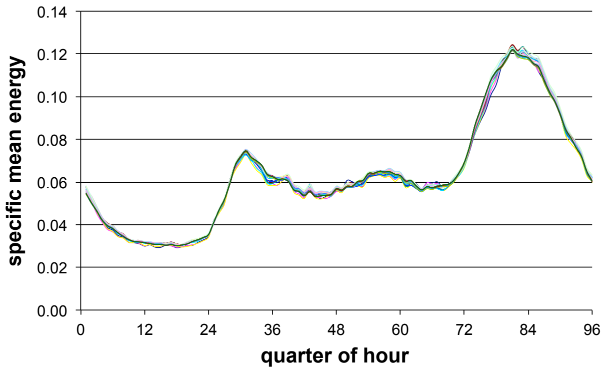


Fig. 15. Specific mean energy of the load curve powers, for different groups of aggregate load curves (from 10 to 50 groups with step 10).

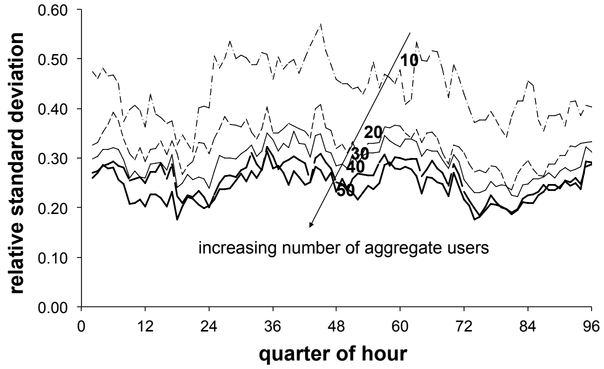


Fig. 16. Relative standard deviation (referring to the average) of the sum of the load curve powers, for different groups of aggregate load curves (from 10 to 50 groups with step 10).

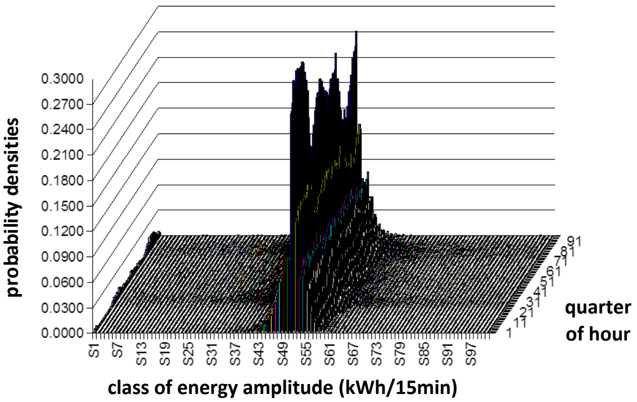


Fig. 17. PDFs of the 15-min energy variations for the set of residential users considered.

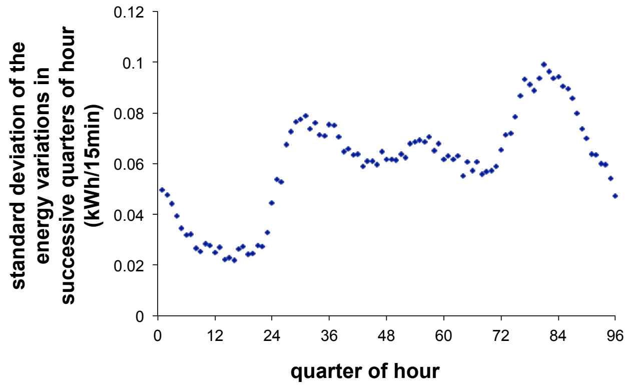


Fig. 18. Standard deviation of the 15-min energy variations for successive quarters of hour.

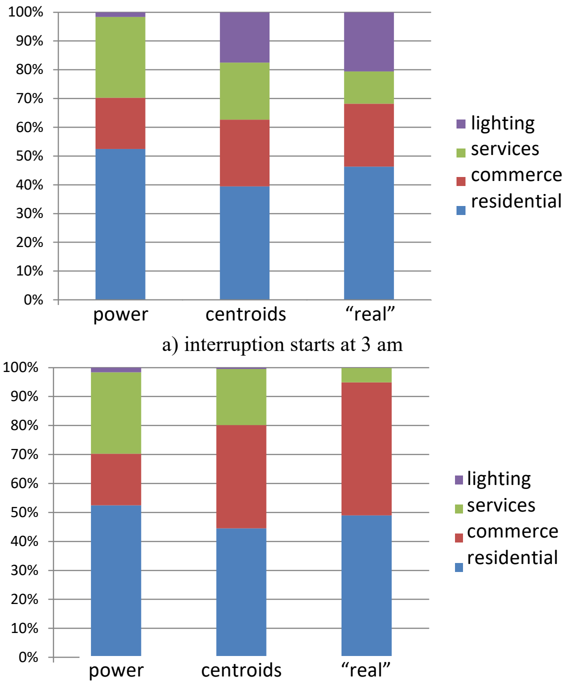
B. *Partitioning of the Energy Not Supplied among Macro-Categories*

The use of load profiles to assess the Energy Not Supplied (*ENS*) following a defined duration of the interruptions that occurs over specified periods of time is practiced in some international contexts. Experiences of this type come from Norway, where since 1999 the *ENS* is calculated through a method based on load profiles for each class of user connected to the supply point. The methodology is described in detail in [41] and includes the use of information regarding the external temperature, the annual energy consumption and the load measured in the hour prior to the interruption (if data are available). A similar approach based on considering the load in the period before the interruption is now used in other jurisdictions (e.g., by the Italian Authority ARERA [42]). However, in the absence of measured data, a *conventional* approach has to be followed. The classical way to obtain a breakdown of the ENS by macro-categories is to use the contract power of the users belonging to the group of interrupted loads. However, in this way no distinction is possible among interruptions occurring at different times of the day, when the composition of the users' demand changes significantly. Conversely, if the conventional load profiles for each macro-category of users are defined, it is possible to estimate the *ENS* of the single macro-category of users considering the integral of the curve that estimates the temporal evolution of the load that would have to be served during the interruption period. The load profiles have to be known before making the evaluation, so that the *ENS* partitioning is indeed conventional and therefore can be accepted by everybody.

Let us consider an example of application to a real group of users connected to the LV network, in which there are four macro-categories of users. The day chosen for the analysis is a weekday (Wednesday 4th April), in which actually there has been no interruption. Hence, a fictitious interruption of one hour starting at different times of the day (chosen at 3 am, 9 am, and 8 pm) is imposed (Fig. 19). The *ENS* breakdown based on the contract power values and on the duration of the interruption results in a single partition to be applied at any time of the day. Alternatively, it is possible to make an *ENS* breakdown on the basis of the load profiles of the macro-categories of users. As an alternative to the *ENS* breakdown by contract power, the use of centroids makes it possible to break down macro-categories closer to the actual use of energy by users (Table III). As a purely indicative example, the *ENS* partitioning in the “real” case (i.e., with the load curves in the absence of the interruption) is added, although of course this partitioning is not relevant for practical purposes if the interruption occurs.

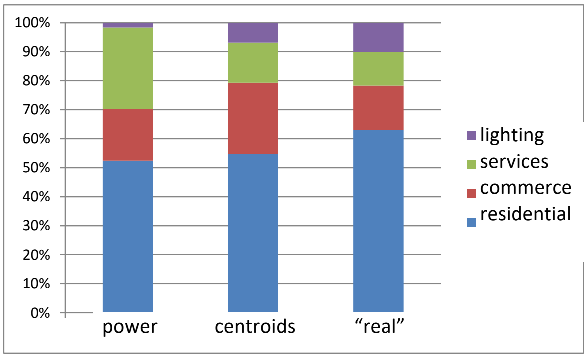
Table III. Example of Partitioning of the Energy Not Served Between Macro-Categories (Wednesday, April 4).

Hour	*Macro-category*	*Partitioning*		
		power	*centroids*	*“real”*
3 am	Residential	52.5%	39.5%	46.3%
	Commerce	17.8%	23.2%	21.9%
	Other services	28.1%	19.8%	11.2%
	Lighting	1.6%	17.5%	20.6%
9 am	Residential	52.5%	44.5%	49.0%
	Commerce	17.8%	35.6%	45.9%
	Other services	28.1%	19.3%	5.0%
	Lighting	1.6%	0.5%	0.1%
8 pm	Residential	52.5%	54.7%	63.1%
	Commerce	17.8%	24.6%	15.3%
	Other services	28.1%	13.9%	11.5%
	Lighting	1.6%	6.8%	10.1%



a) interruption starts at 3 am

b) interruption starts at 9 am



c) interruption starts at 8 pm

Fig. 19. *ENS* partitioning among macro-categories for a 1-hour interruption on Wednesday 4th April.

VI. Conclusions

The SHAPE project has produced converging scientific results in the development of a platform that implements innovative data analytics techniques applied to load curves to obtain customer classification and load profiles. SHAPE is the software prototype for advanced descriptive and predictive analysis of hourly consumption data sourced from smart meters. The results of applying the procedures incorporated in SHAPE lead to a number of benefits for the distribution company. Besides easy visualisation of the load curves at different levels of territorial aggregation, these benefits include better knowledge of the typical load profiles of macro-categories of users at the national level, better knowledge of the load curves for consistent time periods that extend and make more specific the classical concept of seasonality, and automatic classification of the customers on the basis of their actual consumption shapes. Further benefits include the reduction of costs related to the execution of fraud checks, by carrying out targeted campaigns, with respect to the performance of broad-spectrum or sample campaigns, as well as information on the nature of the consumption at the low-voltage level on the national territory, useful to support the National Authority in the definition of dedicated tariffs. The effective deployment of the information coming from the smart metering system paves the way to develop new business strategies and innovative services.

By associating the information provided by the SHAPE platform with the information on the distribution network structure and loads served, a number of benefits may occur for the technical management of the networks. For example, the expected 15-min load can be used to achieve better network reliability through the rationalization of maintenance procedures (e.g., identification of sections of the network most stressed for the same nominal load, and estimation of the time for which a given line has exceeded its thermal limit, with possible costs avoided during maintenance scheduling).

Possible limitations refer to the scalability of some algorithms developed to implement effective clustering procedures, which could be limited in terms of computation speed or memory requirement. However, this is not necessarily a binding constraint, as the maximum number of load curves to be used in clustering procedures is in any case limited by the fact that, due to preliminary macrocategorisation or territorial partitioning, clustering with millions of curves could be practically not necessary.

Further developments include the refinement and update of the procedures implemented, also on the basis of the results obtained by analysing the continuous flow of data that makes new information available, as well as the inclusion in the study of further types of loads such as electric vehicles, other distributed energy resources and storage systems, which are marking the energy transition towards a progressively increasing electrification of the final uses.

At the date of writing the paper, the SHAPE Web portal is not actually deployed in the Company's processes, as e-distribuzione, when it comes to data analytics, is currently focusing on other prediction problems that are particularly addressed by Italian Regulator.

Acknowledgments

The authors wish to thank Mario Ettorre, Lorenzo Gallucci and Massimiliano Ruffolo for their insights and implementation of the procedures.

References

[1] European Smart Grids Task Force - Expert Group 3, *Demand Side Flexibility - Perceived barriers and proposed recommendations*, Final Report, April 2019.

[2] D. Labate, P. Giubbini, G. Chicco, and M. Ettorre, "SHAPE: A new Business Analytics web platform for getting insights on electrical load patterns," *CIRED Workshop*, Rome, Italy, 11–12 June 2014, paper 0354.

[3] D. Labate, P. Giubbini, G. Chicco, and F. Piglione, "SHAPE: The load prediction and non-technical losses modules," *Proc. CIRED 2015*, Lyon, France, 15–18 June 2015, paper 1087.

[4] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Customer Characterisation Options for Improving the Tariff Offer," *IEEE Trans. on Power Systems*, vol. 18, no. 1, pp. 381–387, 2003.

[5] 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 Statistical Society, Part IV*, pp.558–606, 1934.

[6] E. Bompard, E. Carpaneto, G. Chicco, R. Napoli, F. Piglione, P. Postolache and M. Scutariu, "Stratified sampling of the electricity customers for setting up a load profile survey," *Proc. RIMAPS 2000*, Funchal, Madeira, Portugal, September 25–28, 2000, paper RUR-017.

[7] K.-H. Kim, H.-S. Youn, and Y.-C. Kang, "Short-term load forecasting for special days in anomalous load conditions using neural networks and fuzzy inference method," *IEEE Trans. on Power Systems*, vol. 15, no. 2, pp. 559–565, 2000.

[8] G. Chicco, R. Napoli, and F. Piglione, "Load pattern clustering for short-term load forecasting of anomalous days," *Proc. IEEE Porto Power Tech*, vol. 2, 2001.

[9] M.Q. Raza, M. Nadarajah, J. Li, and K.Y. Lee, "Multivariate Ensemble Forecast Framework for Demand Prediction of Anomalous Days," *IEEE Trans. on Sustainable Energy*, in press.

[10] D.F. Rogers and G.G. Polak, "Optimal clustering of time periods for electricity demand-side management," *IEEE Trans. on Power Systems*, vol. 28, no. 4, pp. 3842–3851, 2013.

[11] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, 2012.

[12] A. Notaristefano, G. Chicco and F. Piglione, "Data size reduction with symbolic aggregate approximation for electrical load pattern grouping," *IET Gener. Transm. Distrib*., vol. 7, no. 2, pp. 108–117, 2013.

[13] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Trans. on Power Systems*, vol. 21, no. 2, 933–940, 2006.

[14] X. Li, C. Bowers, and T. Schnier, "Classification of energy consumption in buildings with outlier detection," *IEEE Trans. on Industrial Electronics*, vol. 57, no. 11, pp. 3639–3644, 2010.

[15] S.V. Verdu, M.O. Garcia, C. Senabre, A.G. Marin, and F.J.G. Franco, "Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps," *IEEE Trans. on Power Systems*, vol. 21, no. 4, pp. 1672–1682, 2006.

[16] E. Carpaneto, G. Chicco, R. Napoli, and M. Scutariu, "Electricity customer classification using frequency-domain load pattern data," *Electrical Power & Energy Systems*, vol. 28, no. 1, pp. 13–20, 2006.

[17] M. Petrescu and M. Scutariu, "Load Diagram Characterisation by Means of Wavelet Packet Transform," *Proc. 2nd Balkan Power Conference*, Belgrade, Yugoslavia, 19–21 June 2002, pp. 15–19.

[18] G. Chicco, R. Napoli, F. Piglione, M. Scutariu, P. Postolache, and C. Toader, "Emergent electricity customer classification," *IEE Proc. Generation Transmission and Distribution*, vol. 152, no. 2, pp. 164–172, 2005.

[19] V. Figueiredo, F. Rodrigues, Z. Vale, and J.B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Trans. on Power Systems*, vol. 20, no. 2, pp. 596–602, 2005.

[20] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," *Journal of Knowledge and Information Systems*, vol. 3, no. 3, pp. 263–286, 2001.

[21] M.R. Anderberg, *Cluster analysis for applications*, New York: Academic Press, 1973.

[22] Y.-H. Pao and D.J. Sobajic, "Combined use of unsupervised and supervised learning for dynamic security assessment," *IEEE Trans. on Power Systems*, vol. 7, no. 2, pp. 878–884, 1992.

[23] G.J. Tsekouras, N.D. Hatziargyriou and E.N. Dialynas, "Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers," *IEEE Trans. on Power Systems*, vol. 22, no. 3, pp. 1120–1128, 2007.

[24] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

[25] G.B. Huang, Q.Y. Zhu, and C.K. Siew, "Extreme Learning Machine: Theory and Applications," *Neurocomputing*, vol. 70, pp. 489-501, 2006.

[26] S. Ding, H. Zhao, Y. Zhang, X. Xu, and R. Nie, "Extreme learning machine: algorithm, theory and applications," *Artificial Intelligence Review*, Springer, 2013.

[27] G.B. Huang, Q.Y. Zhu and C.K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," *Proc. IEEE International Joint Conference on Neural Networks*, Budapest, Hungary, vol. 2, 25–29 July 2004.

[28] T.M. Cover, "Geometrical and Statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. on Electronic Computers*, vol. EC-14, pp. 326–334, 1965.

[29] Y.H. Pao, G.H. Park and D. Sobajic, "Learning and generalization characteristics of the random vector functional-link net", *Neurocomputing*, vol. 6, 1994, pp. 163–180.

[30] H. Te Braake and G. Van Straten, Random activation weight neural net (RAWN) for fast non-iterative training, *Engineering Applications of Artificial Intelligence*, vol. 8, no. 1, pp. 71–80, 1995.

[31] P. Guo and M.R. Lyu, "Pseudoinverse learning algorithm for feedforward neural networks," in N. Mastorakis (ed.) *Advances in Neural Networks and Applications*, World Scientific Engineering Society, 2001.

[32] R. Zhang, Z.Y. Dong, Y. Xu, K. Meng, and K.P. Wong, Short-term load forecasting of Australian National Electricity Market by an ensemble model of extreme learning machine, *IET Generation, Transmission & Distribution*, vol. 7, no. 4, pp. 391–397, 2013.

[33] S. Li, L. Goel, and P. Wang, "An ensemble approach for short-term load forecasting by extreme learning machine," *Applied Energy*, vol. 170, pp. 22–29, 2016.

[34] Y. Lin, H. Luo, D. Wang, H. Guo, and K. Zhu, "An Ensemble Model Based on Machine Learning Methods and Data Preprocessing for Short-Term Electric Load Forecasting," *Energies*, no. 10, art. 1186, 2017.

[35] P.J. Werbos, Beyond regression: new tools for prediction and analysis in the behavioural sciences, Ph.D. Thesis, Harvard University, Cambridge, MA, August 1974.

[36] G.A. Darbellay and M. Slama, "Forecasting the short-term demand for electricity: Do neural networks stand a better chance?," *International Journal of Forecasting*, vol. 16, no. 1, pp. 71–83, 2000.

[37] G. Chicco and A.G. Rusu, "Predictability concepts applied to wind and electrical load time series," *Acta Electrotehnica (Special issue)*, vol. 47, no. 4, pp. 135–140, 2006.

[38] I.A. Sajjad, G. Chicco, and R. Napoli, "Definitions of Demand Flexibility for Aggregate Residential Loads," *IEEE Trans. on Smart Grid*, vol. 7, no. 6, pp. 2633–2643, 2016.

[39] S. Haben, C, Singleton and P. Grindrod, "Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data," *IEEE Trans. on Smart Grid*, vol. 7, no. 1, pp. 136–144, 2016.

[40] E. Carpaneto and G. Chicco, "Probabilistic characterisation of the aggregated residential load patterns", *IET Generation, Transmission and Distribution*, vol. 2, no. 3, pp. 373–382, 2008.

[41] J. Heggset, G.H. Kjolle, F. Trengereid, and H.O. Ween, "Quality of supply in the deregulated Norwegian power system," *Proc. 2001 IEEE Porto Power Tech*, Porto, Portugal, 2001.

[42] Italian Regulatory Authority for Energy, Networks and Environment (ARERA), web site https://www.arera.it/it/inglese/index.htm (accessed 16.12.2019).