Comparison and Clustering Analysis of the Daily Electrical Load in Eight European Countries

Pietro Ferraro, Emanuele Crisostomi, Mauro Tucci and Marco Raugi

Department of Energy, Systems, Territory and Constructions Engineering - University of Pisa, L.go Lucio Lazzarino 1, 56122, Pisa, Italy

Abstract

This paper illustrates and compares the ability of several clustering algorithms to correctly associate a given aggregate daily electrical load curve with its corresponding day of the week. In particular, popular clustering algorithms like the Fuzzy c-Means, Spectral Clustering and Expectation Maximization are compared, and it is shown that the best results are obtained if the daily data are compressed with respect to a single feature, namely the so-called “Morning Slope”. Such a feature-based clustering appears to outperform the clustering results obtained upon using other classic features, and also with respect to using other conventional compression methods, such as the Principal Component Analysis, in all the examined European countries. This result is particularly interesting, as this feature provides a direct physical interpretation that can be used to obtain insights on the structure of the daily load profiles.

Keywords: *Clustering*, *Daily Load*, *Expectation Maximization*, *Spectral Clustering*, *Fuzzy c-Means*

1. Introduction

The recent increase of non dispatchable energy sources, that inject power into the grid in a non predictable way, has led to several stability issues in the electrical grid and it is one of the reasons to carefully study and analyse the electrical load consumption data: in order to maintain the balance between produced and consumed energy, several plants (e.g., thermoelectric power plants) are continuously maintained switched on at a low level, as a backup to match the energy demand, if needed (i.e., if they were switched off, then it would take an exceedingly long time before they could effectively be used as a backup). Such an operation is very expensive, especially as it might be very rare that they are used in practice. In this framework, it becomes of primary importance to be able to fully grasp the behaviour of the electrical load: to understand in a clear and quantitative way which parameters characterize the daily load profiles, thus to be able to predict, in an accurate manner, the electrical demand. This will provide useful insights to the energy suppliers in order to schedule the operation of power plants in a more efficient, and economically convenient way; moreover the recent political scenario is shifting toward a stronger deregulation of the energy market: this will allow in a near future the single user to buy energy at a variable price, thus making even more important to have a clear understanding of the electrical load behaviour, for economical reason.

In order to fully characterize the electrical load, many mathematical methods have been employed during the past few decades (Fourier or Wavelet analysis to make an example) resulting in a very deep knowledge of the subject by the scientific community. Despite this, such analyses often employ techniques whose results due to their mathematical content, might prove difficult to be interpreted by a non technical audience; for this reason it appears useful to employ methods whose results have a clear physical meaning that make them understandable by anyone lacking depth of knowledge in the subject at hand. In this perspective, a parameter that plays a crucial role in analyzing the daily electrical load is the day of the week: different behaviours can be observed in working days (e.g. Mondays to Fridays), preholidays (e.g. Saturdays and days prior to festivities) and holidays (e.g. weekends, Christmas and other festivities), while similar trends can be obtained comparing days that belong to the same set. From these considerations, clustering algorithms, due to their ability to find common patterns and due to their natural physical interpretation, appear as optimal candidates.

1.1. Clustering and Classification

Classification and clustering of time series signals is an important area of research in several fields. Clustering refers to the ability to aggregate similar objects together, in groups called *clusters*. The idea of similarity is fundamentally, a human one: it is not trivial to define, in a rigorous way, what *similar* means. Moreover, even if a definition could be found, it would not be an unambiguous one, since it would still depend on the metric used to compare the data. There are numerous motivations to group objects into clusters, an in depth analysis can be found in [1]; in this paper the reasons to use clustering algorithms are

- they have a good predictive power;
- they allow to compress the data into a reduced number of informations.

These properties leads to a more efficient description of the data, which improves the ability to choose the actions to take in specific situations.

1.2. State of the Art

The use of clustering techniques to analyse the electrical load data is a consolidated area: in [2] clustering algorithms are used in order to analyze and divide large electricity customers into classes to estimate their typical days and their representative daily load profiles, in [3] clustering techniques are used to propose an annual framework for optimal price offering. For this purpose, load profiles of customers are used as well as their consumption patterns. In [4] a new clustering algorithm for load profiling is proposed, based on billing data.In [5] an analysis of the load profiles of a representative sample of Spanish residential users is performed by using dynamic clustering (i.e., dynamic in the sense that the load profiles are interpreted as a time series database). A similar attempt has been previously performed, in [6]: the paper shows how to classify electrical customers, in particular, 234 non residential customers in Italy connected to the Medium Voltage (MV) distribution system, using different clustering algorithms; the provided results are not interpreted. In [7], a building in a university campus in Greece is analyzed through a clustering analysis. In some references, clustering analysis is mainly performed as a basis for load forecast. Among others, we remind references [8]-[11]. In particular, in [12], the authors point out that clustering several consumers can lead to an increase in forecasting accuracy. Other related works include [13] in which clustering algorithms are employed to study the electrical load profile and for peak load assessment; in [14] an initial set of centroids, defined by a user defined centroid model, is used to identify load patterns; in [15] the residential electrical load is modelled using mixture model clustering and Markov models; finally in [16] the authors propose a neuro fuzzy classification methods to monitor the load in a non intrusive fashion.

1.3. Objective of the Paper

In the works cited so far, and in the literature in general, very little was done to identify *aggregate* electrical daily patterns using clustering algorithms. It is clear, intuitively, that during the week different electrical behaviour can be observed and classified; the objective of this paper is to provide a mathematical framework to formally describe how the classification can be performed automatically and up to what extent.

A preliminary analysis along the previous lines is described in [17] where it is shown that the task of classifying the data in two different clusters can be performed exceedingly well, obtaining consistent performances of about 95% for each investigated country. The same algorithms though, do not provide the same outcomes for the three classes case, resulting in very poor performance for what regards a direct approach. In the aforementioned paper it was shown that the results can be improved, ranging from 89% to 94%, using a hierarchical approach; in Section 3 this aspect is further investigated.

In both cases though, these performances were not obtained for the direct raw values, for which the results were extremely poor, but on a set of features extracted from the dataset, considered one by one.

This paper extends the work done in [17] along a number of different lines:

- The comparison is extended to more countries of the European union;
- The preliminary results of [17] are rigorously confirmed by using different clustering algorithms;
- We consider PCA as an alternative well-established compression technique to pre-treat data before running a clustering algorithm, and we compare the results with those obtained in [1] (now [17]) using single features (see later Table 2 for a full list of single features);
- A sensitivity analysis is performed on the most informative feature, showing its reliability and robustness.

This paper shows that the feature called *Morning slope* performs better even than a well established compression method like the Principal Component Analysis (PCA), no matter what clustering algorithm is employed, with the advantage that such a feature preserves a clear physical meaning as it is focused on the load values during specific hours of the day. This result, which to the best of the authors' knowledge has never been noticed before, is rather surprising since it appears to hold in very different countries that present load data that are hardly comparable due to different sizes, latitudes and habits (in France, for example, where electrical energy is also used for heating purposes). This classification, being more accurate than the calendar (as it analyzes directly the load curve), can be used as a preliminary analysis for daily load forecasting algorithms in which a different prediction model is used for each cluster [18]. Moreover, on the basis of this classification, average day profiles have been found (see Section 3 for details) which could help energy suppliers to tailor their tariffs.

1.4. Organization of the Paper

This paper is organized as follows: Section 2 introduces the used database, provides some initial insight on the available data, and shortly illustrates the used clustering algorithms and the adopted performance indices. Section 3 thoroughly compares the hourly load patterns among the considered countries in different cases, Section 4 shows the results and the performances of the clustering algorithms and Section 5 provides a detailed description of the differences between the use of the calendar and our analysis. Finally, Section 6 summarize the paper findings.

2. Background: Electrical Load Data and Clustering Algorithms

2.1. Data Set

The data used in this paper are taken from the electrical load data freely available from the ENTSO-E database[1]. ENTSO-E is the European Network of Transmission Systems Operators for Electricity. The ENTSO-E statistical database includes a range of historical data sets regarding power systems of ENTSO-E member Transmission Systems Operators (TSOs). Following the merging of former TSOs associations in 2009, ENTSO-E has become the single data competence centre of the European electricity transmission systems; in particular, 41 TSOs from 34 countries are members of ENTSO-E. Among other data, the ENTSO-E database provides hourly and monthly consumption *aggregated* data for each country (i.e., the whole national load value is given). Note that according to the EC Regulation no. 1228/2003, the national TSOs are obliged to communicate the data related to the electrical physical flows in transmission systems operators' networks.

In this analysis, the daily data of eight countries, namely, Italy, France, Germany, Belgium, United Kingdom, Ireland, Spain and Denmark, from year 2010 to 2014 are employed. The countries were chosen in order to cover most latitudes in Europe (i.e., from Southern to Northern Europe), different sizes, and different electrical loads.

Figure 2 illustrates the electrical load data in year 2012 in the eight considered countries. It shows that there are many differences among the different electrical loads, apart from the obvious difference of the average magnitude of the load which depends among other things, on the size of the countries. In particular, it is possible to note that:

- The electrical load in France is particularly large in winter days. This is due to the fact that electrical energy is also used for heating, as an alternative to gas which is the conventional energy source for heating in (most) of the other European countries;
- The electrical load is particularly high in Italy in summer days due to air conditioning. The same effect is not as clear in other considered countries, due to the fact that the weather is not as hot as in Italy;
- The shape of the electrical loads in Denmark and UK are very similar, despite there is a scale factor of 10^2.

2.2. Clustering Algorithms

We used clustering algorithms to partition N data points (i.e., daily loads) into K classes, where N and K depend on the application of interest. There are many ways to do so and this paper adopts three of the most popular clustering algorithms:

- Fuzzy c-Means;
- Expectation Maximization;
- Spectral Clustering.

In particular, the number of classes that will be investigated amount to three: direct visual inspection of Figure 1 confirms this hypothesis suggesting that the behaviour on festive days and days that precede them, behave differently than normal working days, from an electrical point of view. In what follows the three clusters will be named

- Weekdays,
- Holidays,
- Preholidays,

in order to compare them with the calendar classification. It could be argued that the aforementioned clusters can be directly obtained by extrapolating them from the calendar, thus making the clustering analysis unnecessary; nevertheless, the electrical behaviour of many days is often not associated with their "calendar behaviour" due to the specific cultural habits of each country (e.g. in Italy, the period of time that goes from the 10th to the 25th of August is normally used to schedule vacancies thus resulting in an Holiday period, from an electrical point of view). In Section 5 this aspect will be further discussed.

2.2.1. Fuzzy c-Means

The Fuzzy c-Means (FCM) algorithms is the fuzzy version of the so called k-Means algorithm. In the k-Means algorithm, K initial points are randomly chosen as initial guesses for the center of each cluster; these points are denoted as centroids. At each iteration the position of the centroids is updated and through an *assignment step* each point is assigned to the nearest centroid. Updating the position of centroids is called the *update step*: their position is changed in order to match the means of the points assigned to each respective cluster. It can be proven that this algorithm always converges to a fixed point [1]. The main characteristic of the k-Means algorithm is that each point assigned to a specific cluster equally belongs to that set, without keeping in consideration the distance from the centroid. If this aspect might not prove an issue in many cases it is possible to make the algorithm softer, and associate each points with a degree to quantify the membership to each cluster. The most famous example of such an algorithm is the FCM, initially proposed by Dunn [19] and then refined by Bezdek [20]. It can be shown that basic Fuzzy c-Means algorithms correspond to maximum-likelihood algorithms for fitting a mixture of Gaussians to data [1].

In particular, the algorithm performs the following steps:

- Choose a number K of clusters;
- Randomly assign K centroids;

[1] `https://www.entsoe.eu/data/data-portal/consumption/Pages/default.aspx`

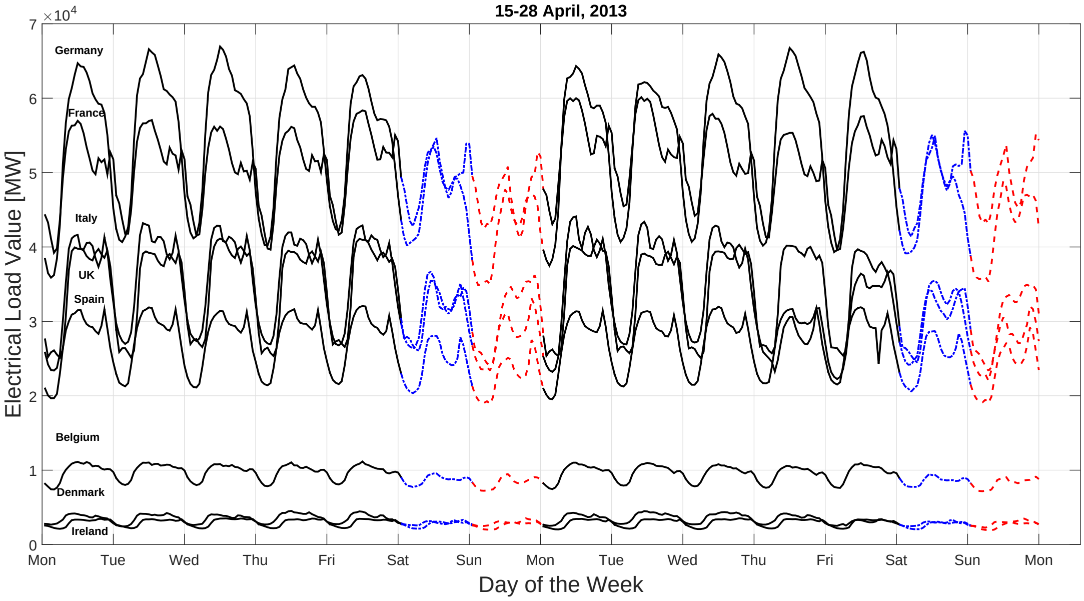

Figure 1: Daily load of the eight countries from April the 15th to April the 28th, 2013. Daily patterns (i.e., weekdays, holidays and preholidays) are clearly visible; the load is typically lower on Sundays and is higher during the weekdays. Some intermediate behaviour is sometimes obtained on Saturdays. The black continuous line represents weekdays, the blue dot dashed line represents preholidays and the dashed red line represents holidays

- Calculate the fuzzy membership value for each point to each cluster (i.e., how much each point belongs to a certain cluster);
- Compute the new centroids on the basis of the fuzzy membership values;
- Repeat step 3 and 4 until convergence.

In particular being interested in finding three hard partitions (namely, weekdays, holidays and preholidays) a basic defuzzyfication step is employed, after running the FCM algorithm, not exploiting the membership degree. In general the FCM and the k-Means perform differently [1].

2.2.2. *Expectation Maximization*

Let $\mathbf{y}$ be a vector of length d (i.e. $\mathbf{y} \in \mathbb{R}^d$), whose value depends on a vector of parameters θ that we wish to estimate. A common way to estimate it is by maximizing the so called maximum log-likelihood. Formally, we want to find $\hat{\theta}$ such that

$$\hat{\theta} = \arg\max_{\theta} log(p(\mathbf{y}|\theta)) \tag{1}$$

where $p(\mathbf{y}|\theta)$ is the probability density function of the random variable $\mathbf{y}$ given the parameter θ.

The EM algorithm solves this problem with the following steps:

- Find an initial guess for the parameters, θ_0;
- Calculate the expected value of the log-likelihood function with the current estimate of the parameters θ_t (E-step);
- Find the parameters that maximize the expected value, calculated in the E-step (M-step);
- Repeat step 2 and 3 until convergence.

This iterative procedure converges to an estimate of the maximum log-likelihood with a monotonically decreasing error. It is worth to point out that due to the possibility of having many local minima, the algorithm needs to be started multiple times in order to obtain an acceptable value. The EM algorithm has been used on a wide spectrum of applications; here, it is applied in order to solve a Gaussian Mixture Modelling problem, called EM clustering: given 3 Gaussians, i.e. the number of clusters that are going to be identified, the algorithm finds the covariance matrices and the means (i.e., the parameter θ) that best fit the data at hand. In order to simplify the problem, it is assumed that each element is generated by only one Gaussian at a time, meaning that each Gaussian corresponds to a cluster

It is interesting to point out that the EM clustering can be previously explained as a generalization of the Fuzzy c-Means algorithm; the latter is, in fact, as pointed out previously, equivalent to a maximum likelihood problem in which each Gaussian component, of the mixture used to fit the data, is spherically shaped, [1].

2.2.3. *Spectral Clustering*

Given a data set of N points, $\chi = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\}$ and a symmetric similarity matrix S, each element being $s_{ij} = ||\mathbf{x_i}\mathbf{x_j}||_S$, with some metric $||\cdot||_S$, it is possible to consider the directed graph $\Lambda(V_\chi, E)$ associated with all these elements. V_χ is the set of vertices, whose elements correspond to the elements of χ

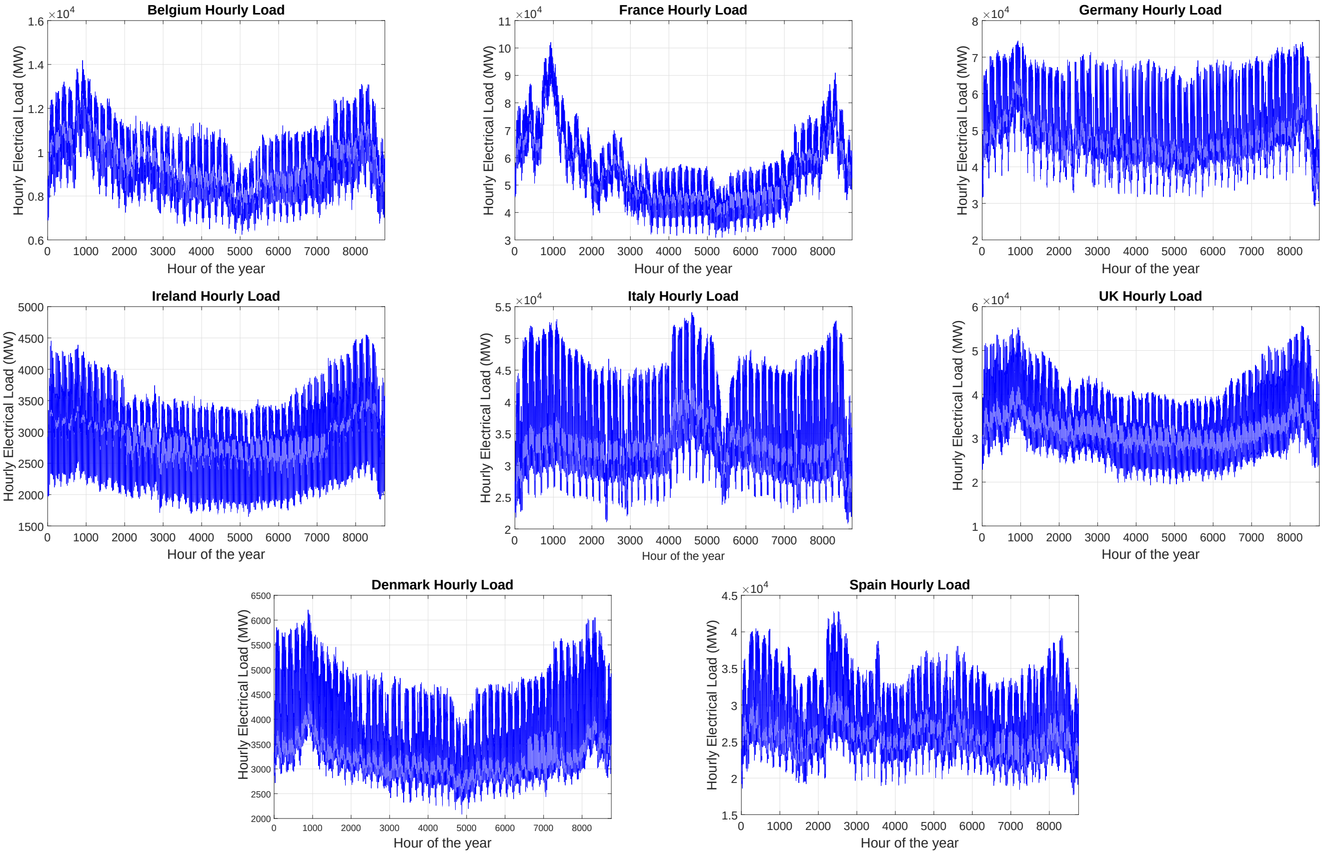

Figure 2: Hourly electrical load in the different considered European countries in the whole year 2012.

and E is the set of edges, whose elements e_{ij} connect the vertices of V_χ. In any directed graph the symmetrical normalized Laplacian matrix is defined as

$$L_{sym} = I - D^{-1/2} W D^{-1/2} \tag{2}$$

W, is the adjacency matrix of the graph, whose values are the elements of E and D is its degree matrix, defined as follows

$$W = (e_{ij})_{i,j=1,\dots,N}, \tag{3}$$

$$d_i = \sum_{j=1}^{N} w_{ij},$$

$$D = diag(d_1, d_2, \dots, d_N). \tag{4}$$

If $w_{ij} = 0$, then $\mathbf{x}_i$ and $\mathbf{x}_j$ are not connected by any edge. For more details on Equation (2) and its interpretation the reader may refer to, among others, [22]. The key idea behind the Spectral Clustering is that the eigenvalues of (2) are all real (since (2) is a symmetric matrix) and it is possible to prove that at least one of them is equal to zero. A further result is that in any graph the algebraic multiplicity of the zero eigenvalue of its Laplacian matrix is equal to the number of disconnected subgraphs [22]; thus, since the vertices V_χ are composed by the data χ and (2) is obtained considering the adjacency matrix W, any number of disconnected subgraphs of $\Lambda(V_\chi, E)$ represent subsets of χ that are totally dissimilar between each other, i.e. the clusters we want to identify.

2.3. *Performance Index*

As pointed out in Section I, the aim of this paper is to test the ability of the previously described clustering algorithms, to automatically classify daily profiles as belonging to one of the three classes that have been identified, by visual inspection, in Figure 1. In particular, we are interested in comparing the results that are obtained in each of the eight selected countries using different compression procedures. Then, the performance index corresponds to the number of days belonging to a given cluster that have been correctly classified divided by the total number of days considered. The “correct” classification is taken from the calendar in order to have a comparison tool, this aspect will be further discussed in section 5. At this regard, note that it was necessary consider a different set of festive days for each country.

3. Methodology: Improving Clustering Algorithms via Feature Based Analysis

The dataset was split into a training and a validation set, accounting respectively for 80% and 20% of the total number of data; this is obtained considering the years ranging from 2010 to 2013 for the training procedure, while the year 2014 is used for validation. After running each algorithm on the training dataset, the three regions in which the points were divided are identified and then the data of the validation set are assigned to the clusters found for the training dataset, based on their values. For the three algorithms the methods employed in order to identify the regions are listed in Table 1.

Table 1: Methods to identify the three regions in which the dataset is divided

Clustering Algorithm	Identifying method
Fuzzy c-Means	The regions are identified by using centroids and the furthest point of each cluster in the training set: the first one identifies the center of each set, the latter identifies its radius
Expectation Maximization	The regions are identified by using the three Gaussian distributions
Spectral Clustering	The regions are identified considering, for each point x, of the validation set, the S nearest points belonging to the training set. The point x is thus assigned to the region to which the majority of the S aforementioned points belong. S was set to 21, by a trial and error procedure.

For what regards the clustering purposes, in [17] it was shown that trying to separate the data directly into three clusters performs quite poorly and that, on the other hand, it is possible to obtain percentages up to 95% for each country separating the data into two clusters; for this reason we decided to tackle the problem with a two step hierarchical approach:

- firstly the data are clustered into two sets, the holidays and the preholidays into one and the weekdays into the other;
- secondly the set containing the holidays and the preholidays is further divided into two clusters, thus obtaining the three sets we are interested in.

In what follows the compression criteria employed are presented and the results discussed, comparing the performance obtained for each country.

3.1. PCA based Clustering

The Principal Component Analysis (PCA) is a well known compression method in which the dataset is multiplied by an orthogonal operator in order to obtain a new and smaller set of variables that are as informative as possible with respect of the original data [23]. This operator is defined in such a way that the first principal component has the largest possible variance and each succeeding one in turn has the highest possible variance, under the constraint that it is orthogonal to the preceding ones.

Table 2: List of interesting features

Feature	Definition
Daily Mean (Mean)	Mean of daily load values
Daily Variance (Var)	Variance of daily load values
Min-Max (MM)	Difference between the maximum and the minimum value of the daily load
Max Peak (MP)	Maximum value of the daily load
Morning Slope (MS)	Difference between the load value at 10.00 am and at 06.00 am
Partial Daily Mean (PDM)	Mean of daily load values between 11.00 am and 08.00 pm
Partial Daily Variance (PDV)	Variance of the daily load values between 11.00 am and 08.00 pm
Partial Min-diff (PMD)	Difference between the average load and the minimum load between 11.00 am and 08.00 pm
Early Afternoon (EA)	Difference between the load value at 11.00 am and at 03.00 pm
FFT Peak (FFTP)	Maximum of the absolute values of the Fast Fourier Transform of the daily load values

In this paper the PCA compression is performed on the normalized data correlation matrix and we picked a number of components for each country based on $\sigma_i^2 \geq \rho\sigma_1^2$, with σ_i being the standard deviation of the ith component. The value ρ was set to 0.025 that is, in our experience, a good trade off value to remove the noisy components of the dataset. With the aforementioned criteria the number of Principal Components for each country was set to three, with the exception of Belgium, France and Spain for which the number of Components was set, respectively, two, two and four.

Figure 3 shows the Principal components of the Germany dataset; it is clear, from visual inspection, that after the third component, the variance contribution becomes negligible.

3.2. Feature based Clustering

As shown in [17], the use of features in place of raw data drastically improves the performance of the clustering algorithms for the purpose at hand. On the basis of such results we consider here the features listed in Table 2.

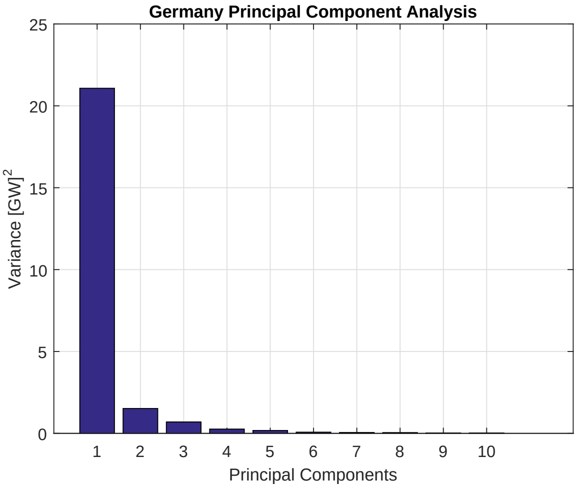

Figure 3: The first 10 principal Components of the Germany dataset

Some of them are conventional options in data analysis, others are known to be interesting for the specific application of interest. Each of them was evaluated separately from the other ones, clustering the data using one feature at a time. Even though, as pointed out in the introduction, the *Morning Slope* is the best performing feature, it is interesting to perform a comparison with the other ones, to show that the informations extracted do not allow, in general, to obtain satisfactory performances in a consistent way for all the eight considered countries.

4. Validation Results

Figure 4 and Tables 3 and 4 give the results for each clustering algorithm performed on the features proposed in Table 2 and on the PCA compressed data. We first list the training results performed on the years 2010-2013 and then the validation ones, obtained in year 2014. Due to space limits and for readability purposes, the complete performance Figure is given only for the Fuzzy c-Means algorithm on the training case, while for the others only the results obtained are listed on tables comparing the most informative feature and the PCA results. It is possible to see that the *Morning Slope* outperforms every other features (PCA included) in every country. Slightly worse results are obtained for the case of Spain (even in this case though, the *Morning Slope* is still the most informative feature in every test, consistently with the other countries). It is interesting to stress the fact that a simple feature such as *Morning Slope* outperforms a well established compression method like PCA, in this specific problem no matter what algorithm is employed. The PCA compressed data, on the other hand, exhibits poor performance for what regards the Fuzzy c-Means, while performing better with the Expectation Maximization algorithm; this appears reasonable since, on multiple dimensional data (such as the PCA), unless the variance on each axis is the same (i.e., the sets are spherical), or the clusters are very distant from each other (so that, despite their shape they are easily separable by spheres), clusters with different shapes and orientations needs to be employed, thus making a simple algorithm such as the Fuzzy c-Means ineffective. It is then interesting to see that the performances of the algorithms are very different between the training and the validation case: in the latter, the results degrade significantly, especially in the Fuzzy c-Means and Expectation Maximization case, while Table 4 shows that the Spectral Clustering performs more consistently, obtaining similar results between the training and the validation case, thus making the Spectral Clustering the best performing algorithm overall (at least, for what regards the *Morning Slope* feature). Lastly, Figure 5 shows a sensitivity analysis for what regards the *Morning slope*: the results are obtained by changing the bounds of the feature by one hour (plus one, minus one) and enumerating the nine possibilities. Figure 5 shows that despite the classical feature (ranging 06:00-10:00 AM) performs better for the majority of the countries, with slight changes for what regards Belgium and Denmark, for Spain and Ireland the results are greatly improved by the change of time intervals. This appears reasonable since the daily habits are different from a country to another thus affecting the hours at which the *Morning Slope* is most effective at classifying each day.

5. Discussion: Load Based Clustering vs Calendar Clustering

The use of the features in order to cluster the day of the year might appear somewhat arbitrary: it could be argued that the calendar would give a performance of 100% and is thus better than any other classifying method. Despite that, the load profile of some days of the year does not follow the typical behaviour of that day, up to the calendar classification (see, for instance the Italian hourly load in Figure 2, where it is obvious that the load is very low in the middle of the figure, due to "summer holidays" that, according to the calendar, are working days by all means). For this reason, in order to better clarify the scope of this research, the main differences between the clusters obtained using the calendar and the clusters obtained using the proposed features, are listed below:

- The calendar classification can not be considered 100% accurate in order to capture the actual behaviour of a daily load. The reason is that some days of the year, though belonging to a certain class just from the point of view of the calendar, still, in every day life, they correspond to another class (i.e., a working day on the calendar might behave like a Holiday or a Preholiday, from an electrical point of view). As an example, as already mentioned, in Italy practically every single office and most industries are closed for two weeks around August 15; thus, all the weekdays appear as "working days" from the calendar, but it is well known that in truth they are Holidays. Obviously, the load follows the real life trend, rather than the calendar. Figure 6 shows that the clustering procedure is consistent with the previous consideration, labeling the days of the

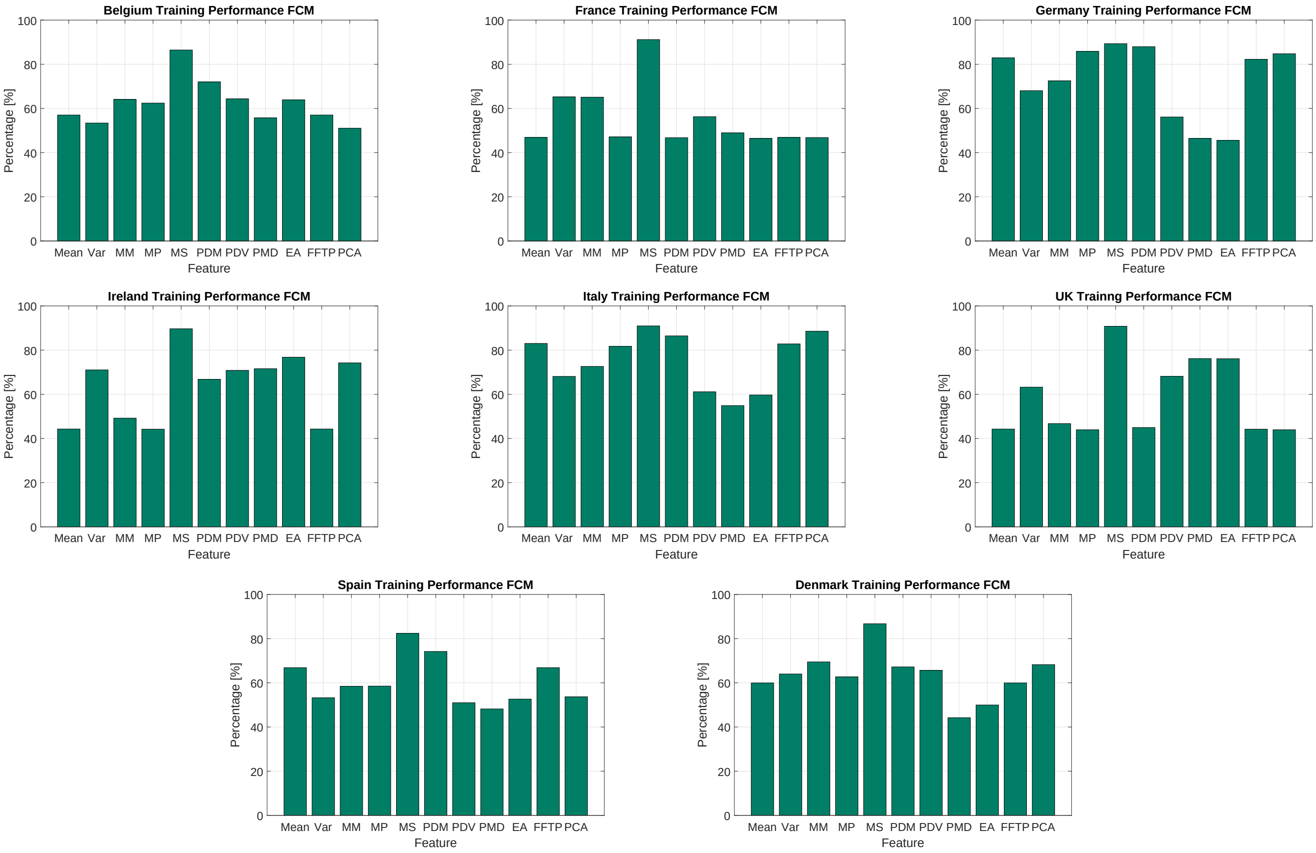

Figure 4: Performance of the Fuzzy c-Means algorithm on the Training set

Table 3: Performance of the Fuzzy c-Means algorithm on the validation set

Validation Performance								
Feature/Country	**Belgium**	**Italy**	**Germany**	**France**	**UK**	**Ireland**	**Denmark**	**Spain**
Morning Slope	**89.863%**	**89.041%**	**92.054%**	**80.273%**	**81.780%**	**73.698%**	**80.547%**	**74.246%**
PCA	53.698%	87.945%	84.657%	38.904%	50.411%	48.767%	47.671%	61.643%

Table 4: Performance of the spectral clustering and expectation maximization algorithm.

Spectral Clustering - Training Set								
Feature/Country	**Belgium**	**Italy**	**Germany**	**France**	**UK**	**Ireland**	**Denmark**	**Spain**
Morning Slope	**87.816%**	**90.759%**	**88.774%**	**89.938%**	**88.843%**	**88.843%**	**84.941%**	**82.203%**
PCA	52.156%	88.569%	86.652%	43.394%	45.653%	47.570%	74.674%	53.456%
Spectral Clustering - Validation Set								
Morning Slope	**85.479%**	**89.863%**	**93.150%**	**89.253%**	**83.287%**	**82.876%**	**81.369%**	**80.767%**
PCA	56.986%	87.123%	84.547%	38.356%	58.356%	53.424%	48.767%	63.287%
Expectation Maximization - Training Set								
Morning Slope	**88.648%**	**92.881%**	**92.072%**	**89.993%**	**87.159%**	**91.863%**	**83.493%**	**85.626%**
PCA	87.542%	89.010%	88.227%	88.090%	86.5161%	85.557%	79.192%	82.612%
Expectation Maximization - Validation Set								
Morning Slope	**89.918%**	**92.013%**	**91.435%**	**80.278%**	**85.789%**	**74.659%**	**79.905%**	**80.137%**
PCA	89.589%	90.137%	80.274%	78.943%	83.561%	72.876%	71.780%	79.437%

weeks around the 15th of August as belonging to the Pre-holiday and Holiday clusters, while the calendar classification is obviously different. As from the Figure the feature based clustering is closer to the daily load curves than the calendar, even simply from visual inspection. Therefore, the only way to increase the matching probability closer

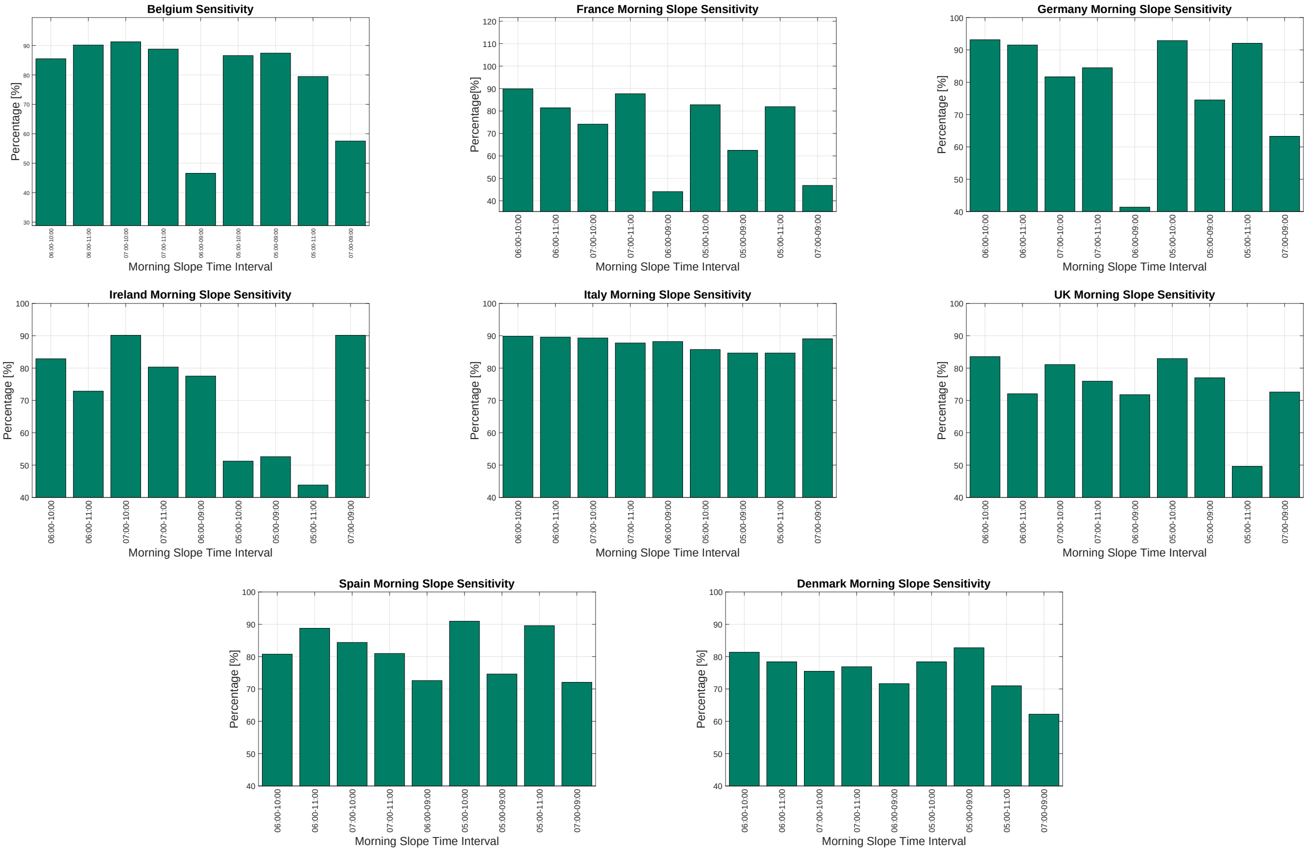

Figure 5: Sensitivity analysis of the *morning slope*. The considered set is the validation one.

to 100 %, is to “redefine“ the calendar according to people’s habits. Since any choice in this direction can be interpreted as arbitrary, and as in any case such operations require some knowledge of the habits of people in different countries, in this paper the calendar is used as a term of comparison. However, one should be aware that higher matching values might be, considering the results shown in figure 6, inconsistent with the true electrical behaviour of each day.

- The classification obtained using the Fuzzy c-Means and the Expectation Maximization provides, by use of the membership function (for the Fuzzy c-Means) and the Gaussian distribution (for the Expectation Maximization) a further information: the extent to which each day belongs to each cluster. The calendar, on the contrary, provides a hard partition, associating each daily load with a single class with membership value 1 and 0 to the other clusters.

- It is well known that clusters have a good predictive power: to forecast the electrical load it is common practice to divide the data into classes and to use a different predictive model for each cluster (or, in this case, a combination of each model, based on the membership values obtained by the Fuzzy c-Means and the Expectation Maximization) in order to improve the accuracy of the predictions [18]. Clearly, in order to obtain accurate results, each day must belong to the “correct” cluster, an information that the calendar does not always provide accurately enough(see Figure 6).

Finally, the validation of the clustering outcomes against the calendar is done purely to compare the abilities of the three algorithms: the fact that the three obtained classes are consistent with the calendar is an a posteriori result due to having chosen the best performing feature for this objective function.

6. Conclusion

As initially anticipated, one of the main advantages of clustering lies in summarizing information in a single data vector (i.e., the centroid of the cluster for what regards the Fuzzy c-Means, the mean of each Gaussian for what regards the Expectation maximization and the eigenmap for the Spectral Clustering); among the considered possibilities the *Morning Slope* emerges as the most informative one, even in comparison with the Principal Component Analysis, obtaining consistently a higher performance for every considered country. According to the previous considerations, the results of the previous analysis can be summarized as follows:

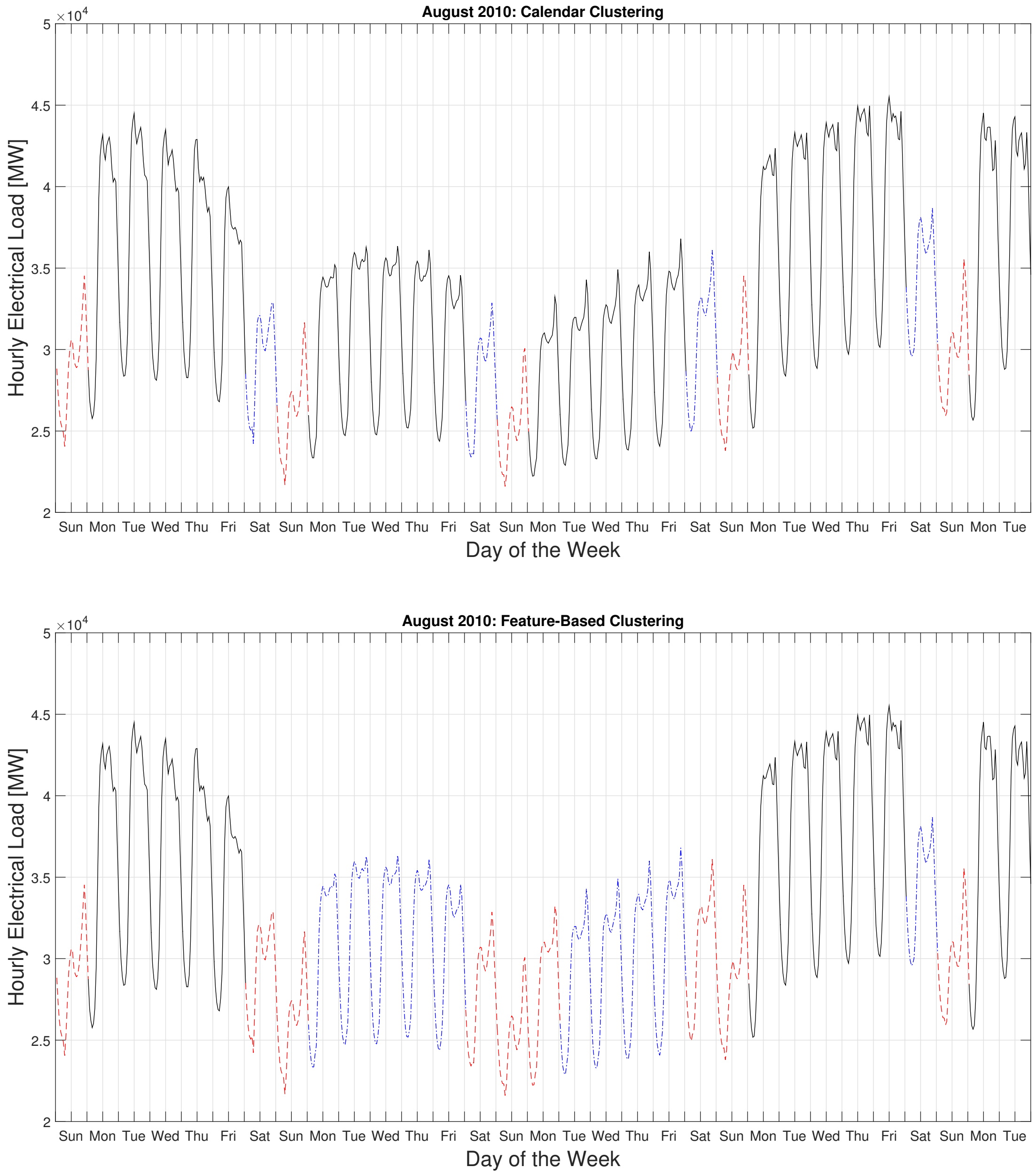

Figure 6: Classification of the days of August 2010. The black continuous line represents Weekdays, the blue dot dashed line represents Preholidays and the dashed red line represents Holidays. The figure above represents the calendar classification, while the figure below represents the classification obtained with the clustering analysis.

- The *Morning Slope* has been identified, as a suitable feature to properly assign the days of each year to three different classes;

- The PCA-based analysis performed poorly with respect to the feature based one showing that the *Morning Slope* is not a trivial parameter (at least, not for the task at hand).

- The sensitivity of the *Morning Slope* has been investigated, showing that each country possesses a slightly different optimal range;

- The obtained results are more consistent with the real life trend than the calendar as they correctly classify “unusual” days;

- Despite the fact that the best performances are achieved using the Spectral clustering algorithm, the use of the Fuzzy c-Means and the Expectation Maximization allows to define the extent to which, each day belong to each cluster, by use of the membership function (for the Fuzzy c-Means) and the Gaussian distribution (for the Expectation Maximization) obtained, thus providing an additional information.

These informations can be used, for instance, to preprocess data for forecasting purposes, a task for which the calendar classification is not always enough. Another possible application is shown on figure 7. Each curve is obtained by considering the average of the hourly load for every day that belong to one of the three clusters. The figure shows that there are major differences in terms of load between the days of each set. It is the authors' belief that this information can be very useful for electrical energy suppliers and retails as an indication of some average profiles to plan optimal scheduling of dispatchable power plants and to tailor ad hoc tariffs to customers accordingly to the day of the week. Again, this information is not obtainable using the calendar alone due to the presence of "unusual" days.

References

[1] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, VI Edition, 2007.

[2] G.J. Tsekouras, P.B. Kotoulas, C.D. Tsirekis, E.N. Dialynas and N.D. Hatziargyriou, *A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers*, Elsevier, Electric Power Systems Research, vol. 78, pp. 1494-1510, 2008.

[3] N. Mahmoudi-Kohan, M. Parsa Moghaddam and M.K. Sheikh-El-Eslami, *An annual framework for clustering-based pricing for an electricity retailer*, Elsevier, Electric Power Systems Research, vol 80, pp. 1042-1048, 2010.

[4] J. Nuno Fidalgo, M. A. Matos and L. Ribeiro, *A new clustering algorithm for load profiling based on billing data*, Elsevier, Electric Power Systems Research, vol. 82, pp. 27-33, 2012.

[5] I. Benítez, A. Quijano, J.-L. Díez and I. Delgado, *Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers*, Elsevier, Electrical Power and Energy Systems, vol. 55, pp. 437-448, 2014.

[6] G. Chicco, R. Napoli and F. Piglione, *Comparisons among clustering techniques for electricity customer classification*, IEEE Transactions on Power Systems, vol. 21, no. 2, pp. 933-940, 2006.

[7] I.P. Panapakidis, T.A. Papadopoulos, G.C. Christoforidis and G.K. Papagiannis, *Analysis of the Electricity Demand Patterns of a Building in a University Campus*, IEEE 12th International Conference on Environment and Electrical Engineering, 2013.

[8] G. Grigoraç, F. Scarlatache and G. Cârţină, *Load estimation for distribution systems using clustering techniques*, IEEE International Conference on Optimization of Electrical and Electronic Equipment, 2012.

[9] Y.J. Xia, Y.H. Yang, F. Ge, J. Su and H. Yu, *Pattern Analysis for Load Forecasting*, IEEE 8th International Conference on Computing Technology and Information Management (ICCM), 2012.

[10] V.S. Kodogiannis and I. Petrounias, *Power Load Forecasting Using Adaptive Fuzzy Inference Neural Networks*, 6th IEEE International Conference on Intelligent Systems, 2012.

[11] I.P. Panapakidis, M.C. Alexiadis, and G.K. Papagiannis, *Load Profiling in the Deregulated Electricity Markets: A Review of the Applications*, 9th IEEE International Conference on the European Energy Market, 2012.

[12] S. Humeau, T.K. Wijaya, M. Vasirani and K. Aberer, *Electricity Load Forecasting for Residential Customers: Exploiting Aggregation and Correlation between Households*, 3rd IFIP Conference on Sustainable Internet and ICT for Sustainability, Palermo, Italy, 2013.

[13] D.D. Sharma and S.N. Singh, *Electrical Load Profile Analysis and Peak Load Assessment using Clustering Technique*, IEEE PES General Meeting, National Harbor, MD, 2014.

[14] G. Chicco, O.-M. Ionel and R. Porumb, *Electrical Load Pattern Grouping Based on Centroid Model With Ant Colony Clustering*, IEEE Transactions on Power Systems, vol. 28, no. 2, pp. 1706-1715, 2013.

[15] W. Labeeuw and G. Deconinck, *Residential Electrical Load Model Based on Mixture Model Clustering and Markov Models*, IEEE Transactions on Industrial Informatics, vol. 9, no. 3, pp. 1561-1569, 2013.

[16] Y.-H. Lin and M.-S. Tsai, *Non Intrusive Load Monitoring by Novel Neuro Fuzzy Classification Considering Uncertainties*, IEEE Transactions on Smart Grid, vol. 5, no. 5, pp. 2376-2384, 2014.

[17] A. K. Tanwar, E. Crisostomi, P.Ferraro, M.Tucci, M. Raugi, G. Giunta, *Clustering Analysis of the Electrical Load in European Countries*, IEEE International Joint Conference on Neural Networks, 2015.

[18] K.Y. Lee, Y.T. Cha and J.H. Park, *Short-term load forecasting using an artificial neural network*, Transactions on Power Systems, Vol. 7, No. 1, pp. 124-132, February 1992.

[19] J.C. Dunn, *A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters*, Journal of Cybernetics, vol. 3, No. 3, 1973.

[20] J.C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York, 1981.

[21] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, Springer, II Edition, 2008.

[22] U. Von Luxburg, *A tutorial on Spectral Clustering*, Statistics and computing, 17.4, pp. 395-416, 2007.

[23] I.T. Jolliffe, *Principal Component Analysis*, Springer series in statistics, II Edition, 2002.

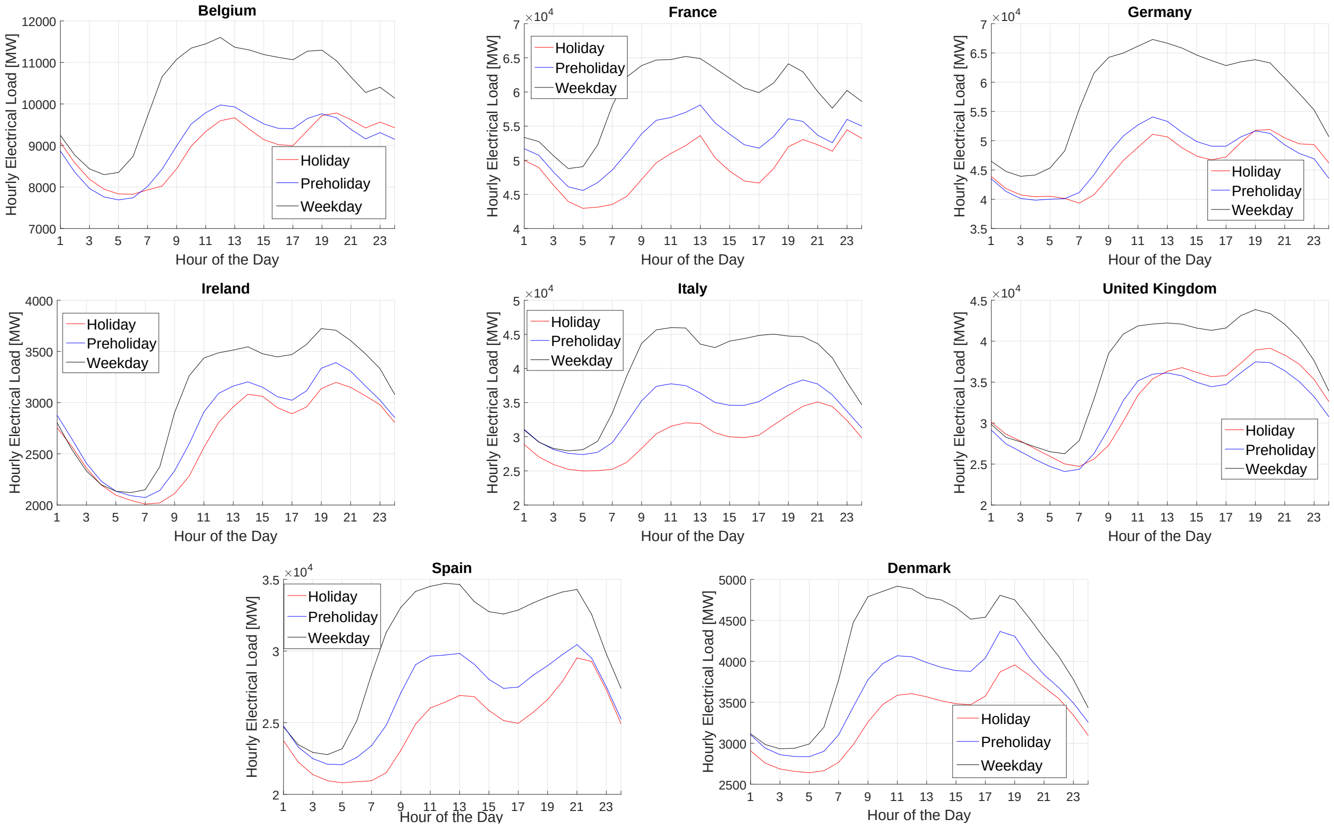

Figure 7: Average day profiles for each country: the curves are obtained considering the average of the hourly load, for every days that belong to one of the three clusters