IDENTIFICATION OF ELECTRICAL ENERGY CONSUMPTION PATTERNS

Vera PEREIRA
EDP Distribuição – Portugal
vera.pereira@edp.pt

Pedro MOUSINHO
EDP Distribuição – Portugal
pedro.mousinho@edp.pt

Luísa JORGE
EDP Distribuição – Portugal
luisa.jorge@edp.pt

ABSTRACT

In the current context of changes in terms of processes and information systems, the development of methodologies to deal with large volumes of information proves to be extremely important. Encouraged by the increasing technological development, which enables the storing and processing of large amounts of data, Data Analysis tools are increasingly present in companies like EDP Distribuição, in order to take full advantage of the available consumption and production data.

INTRODUCTION

Data Analysis techniques make it possible to identify patterns that otherwise would hardly be found and reveal hidden correlations, among other useful information.
In this context, a methodology aiming at segmenting the universe of Primary Substations (PS) according to their patterns of power consumption is presented in this paper.

The database of the study consisted on the active power values of the PS' transformers, as well as on the active power of the Distribution Generation (DG). In both cases, the values were in 15 minutes samples. Because it was the most recent year for which the data was available, the time period analysed was the year 2014.
With these two quantities (active power of PS and active power of DG), the gross load was calculated. In the time period in analysis, there were 384 PS, and so it was easy to see that the amount of data we were dealing in this study was significant.

Therefore, we thought it would be convenient to use a programming software capable of store and manipulate effectively greats amounts of data. We opted to develop the project using 'R' programming language. Being a free software, it provides a wide variety of statistical computing tools and graphical environments.

INITIAL CONSTRAINTS

After a short analysis, several problems were detected concerning the quality of the data being studied. The first one consisted on a time asynchronism between the meters from the PS and the DG. To try to minimize this issue, we used a smoothing method, which consisted in a centered moving average of 5 periods.
Another problem found was the missing data. Because those failures could compromise the results, the days containing missing values were eliminated. Additional problems were found, such as the inexistence of A^- in some meters, the incorrect match between PS and DG and also long reconfiguration periods in the medium voltage (MV) network. Due to the misleading conclusions it could result in, we decided not to consider the data from the PS affected with these problems, as the remaining universe is still significant.

METHODOLOGY

After identifying the problems described previously, and consequent elimination of the PS affected by them, the universe in study changed from 384 to 354 PS.
The methodology developed to identify the power consumption patterns, which allowed the segmentation of the universe of PS, is described below.

As a result of a first analysis, it was perceived that the power consumption behaviours vary according to the season of the year and day type (Business Day, Saturday or Sunday). Therefore, the first step of the methodology consisted in disaggregating the data according to these specifications. Having in mind that our goal was to capture typical behaviours, only the two characteristic months of each season of the year were considered. In other words, in Winter we studied January and February, in Spring, April and May, in Summer, July and August and in Autumn, October and November.

As there are many similarities between the power consumption in days with the same characteristics (for example, the power consumption values in business days of a particular season of the year are very alike between each other), the second step was to determine a unique load profile representative of all the same day types. To do this, we identified and eliminated, using clustering methods, the days with an abnormal power consumption behaviour over the period of study, and calculated the mean of the remaining ones.

The process described on the previous paragraphs was repeated for the several seasons of the year and, and in the end, we obtained a Resulting Load Profile which reflects the annual behaviour of each PS. In Figure 1 we can see the result for a particular PS, called 'Alameda'.

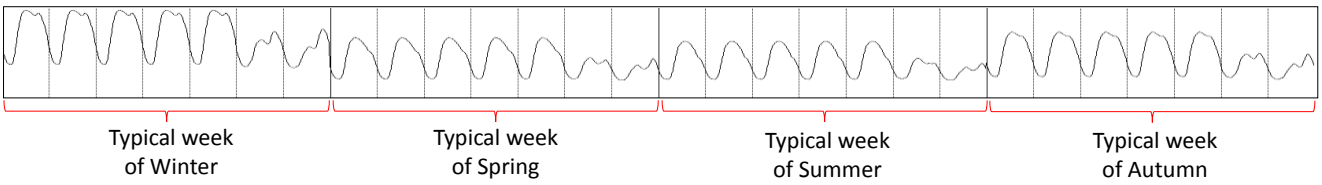

Figure 1- Resulting Load Profile for the PS 'Alameda'.

Naturally, the load profiles that represent the business days should have weight 5 compared to the other days, so that each season of the year is represented by a typical week.
The next step of the methodology consisted in grouping the Resulting Load Profiles in clusters, according to the power consumption patterns. Finally, we intended that each cluster had a representative Typical Load Profile.

Clustering Methods

Figure 2 shows the result of the implementation of the first step of the methodology mentioned previously, applied to the PS 'Aroeira'. In it, we can see the load profile of all its business days in April and May represented simultaneously. As the goal was to have only one representative load profile for all those days, we needed to identify and eliminate the days whose power consumption behaviour was different from the common and calculate the average of the remaining ones. Observing the figure below, we clearly understand that there is a dense area corresponding to the days with a normal behaviour, but also some we want to eliminate.

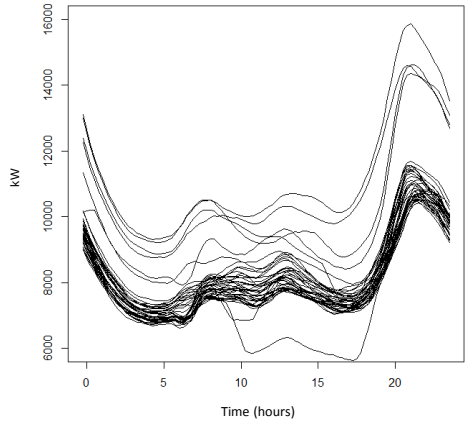

Figure 2 – Load profile of all the business days for the PS 'Aroeira' in April and May.

To automatically identify this dense area we used a Data Mining technique called 'Clustering'. These tools can be described as algorithms that allow us to automatically group objects according to their degree of similarity, such as distance, shape or other.

Density-Based Clustering

Having in mind the goal described above, the clustering technique that best suited our problem was the density-based clustering. The algorithm developed on the programming language 'R' is called 'DBSCAN' (Density-Based Spatial Clustering of Applications with Noise) and intend to group objects according to the density of the area. Therefore, elements with many 'neighbours' (dense area) are put in the same group [3]-[6].
In this initial stage of the work, the elements to group were time series (days) and, as we wanted to measure the proximity between them, the similarity criterion used was the Euclidean distance (1).

$$\delta(M,S) = \sqrt{\sum_{t=1}^{H}(m_t - s_t)^2}\,,\ H = 4 \times 24 \qquad (1)$$

where M and S are time series.

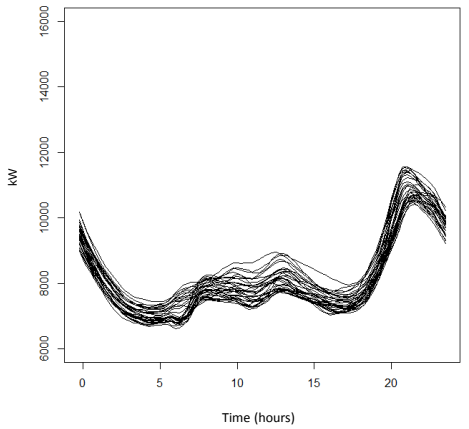

Figure 3 – Result of the application of DBSCAN to the example of Figure 2.

After applying this procedure to all the PS and to the several day types, we obtained a Resulting Load Profile which reflects the annual behavior of each PS. Four of these load profiles are shown in Figure 4.

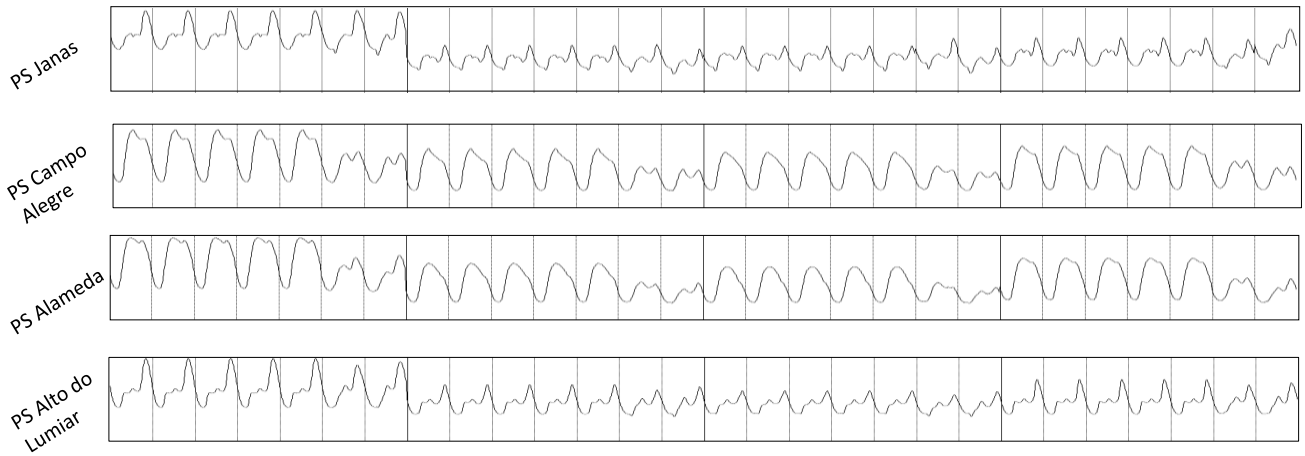

Figure 4 – Example of Resulting Load Profiles.

The next step of the methodology consisted in grouping these Resulting Load Profiles according to their power consumption patterns. Analysing the Figure 4 we understand that, for example, PS 'Janas' and PS 'Alto do Lumiar' have a similar behaviour. In the same way, we can also say that 'Campo Alegre' and 'Alameda' are very much alike in terms of power consumption patterns. As the aim of the methodology was to compare the load profiles in terms of their shape, and not of their average load or amplitude, we needed to standardize the curves.

Hierarchical Clustering

Through the standardization process, the load profiles become comparable and so analysis like the one at the end of last section can be done. To do this automatically, we used once again a clustering technique. This time, the one that seemed to best suit our needs was the hierarchical clustering. In a very brief way, we can say that this algorithm organizes all the elements in a tree (dendrogram), whose leafs represent the PS and the distance between the nots indicates the similarity between them. Thus, similar installations are put in the same sub-trees [3]-[5].

The tree that represents the universe of PS being studied is shown in Figure 5. With a closer look at the first sub-tree, for example, we can see that the PS located in Algarve and Costa Vicentina were put together, reinsuring the intuitive idea we had that these PS, for being highly influenced by the Summer tourism, would have the same power consumption behaviour.

It is important to underline that to do this process, the similarity criterion used considered, not only the proximity of values (Euclidian distance), but also the shape of the load profiles (time correlation). The mathematical expression that shows that relation is:

$$d(x,y) = \Phi[CORT(x,y)]\delta(x,y) \quad (2)$$

In the last equation, the variable that evaluates the shape of the curves ($\Phi[CORT(x,y)]$) measures the proximity between the dynamic behaviors of the time series x and y, by means of the first order temporal correlation, defined by:

$$CORT(x,y) = \frac{\sum_{t=1}(x_{t+1} - x_t)(y_{t+1} - y_t)}{\sqrt{\sum_{t=1}(x_{t+1} - x_t)^2}\sqrt{\sum_{t=1}(y_{t+1} - y_t)^2}} \quad (3)$$

Where $\Phi[u]$ is an adaptive tuning function taking form:

$$\Phi[u] = \frac{2}{1 + e^{ku}} \quad (4)$$

with $k \geq 0$ so that both Φ and k modulate the weight that $CORT(x,y)$ has on $d(x,y)$.

Finally, $\delta(x,y)$ is the Euclidian distance defined in (1).

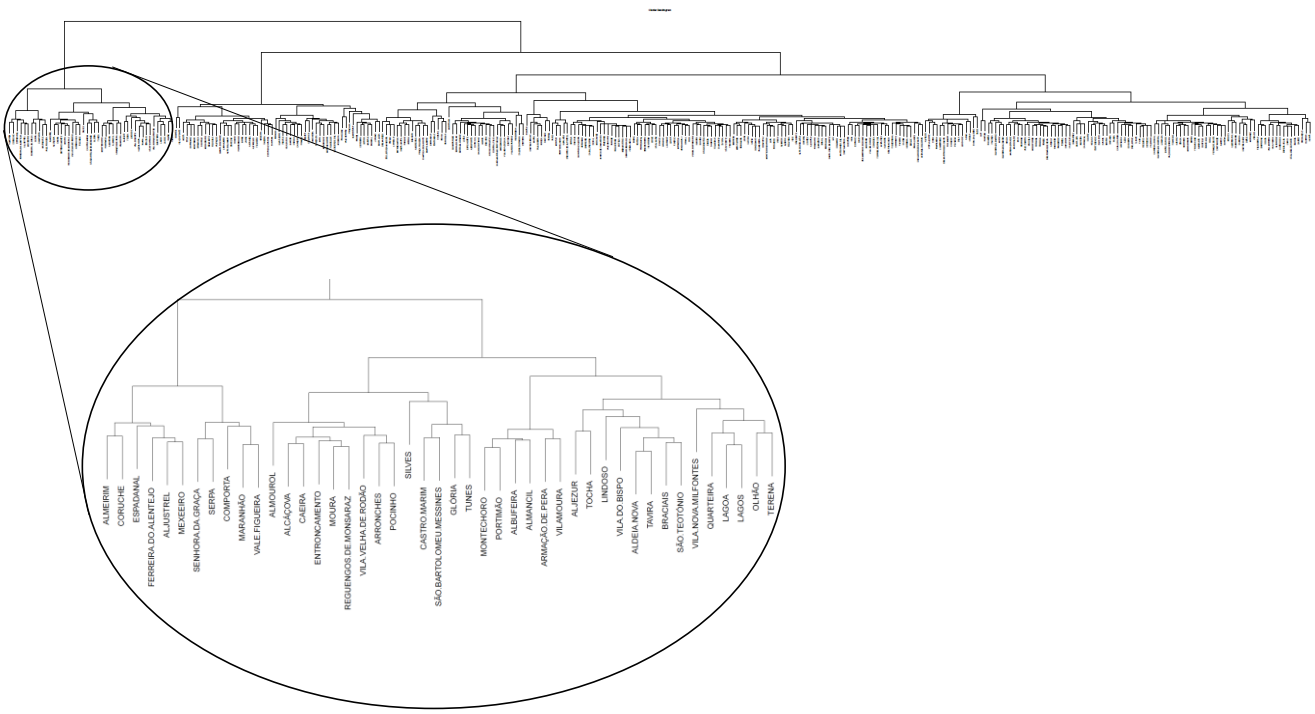

Figure 5 – Hierarchical tree representing the universe of PS in study.

RESULTS

After organizing the PS in a hierarchical tree that reflects the similarities between them in terms of power consumption behaviour, we needed to define the number of groups to be considered. Intuitively, we understand that increasing the number of clusters, the related errors decrease. However, so that the application of the methodology did not involve a very high degree of complexity, which would make it difficult to use, the number of clusters could not be very high. Having that in mind, we needed to define a value which had these two concerns in consideration. After testing several scenarios, we came to the conclusion that the most appropriated number of clusters would be seven. It can be seen in Figure 6 a summary of the obtained results or, in other words, the groups that characterize the behaviour of the MV network, represented by its average load profile, for each season of the year and day type.

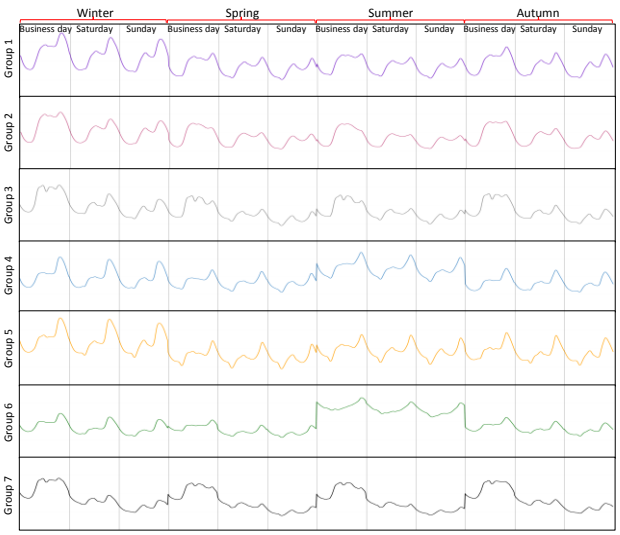

Figure 6 – Summary table with the Typical Load Profiles.

We also complemented the analysis with the geographical location of all the PS in each group, allowing a deeper knowledge of the power consumption behaviours throughout the Portuguese national territory (Figure 7).

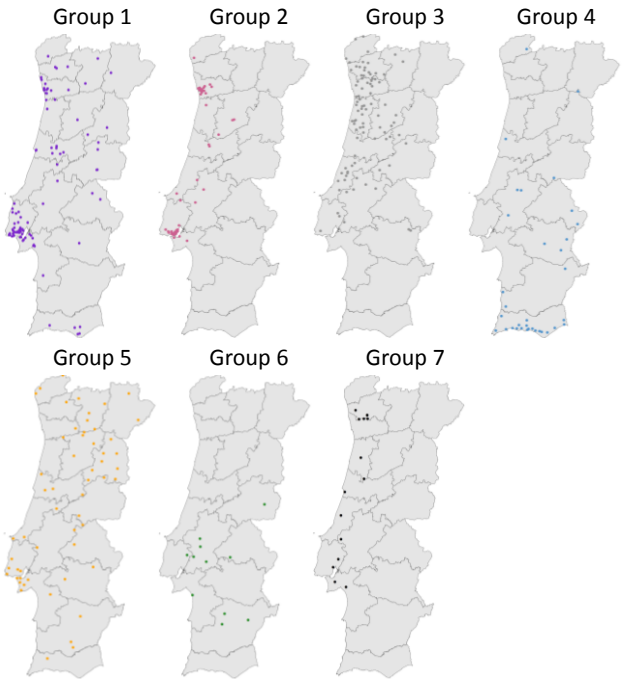

Figure 7 – Geographical Location of the PS in each group.

CONCLUSIONS

This work intended to study large databases through the use of Data Analysis tools. These allow to inspect, filter, transform and model data in order to discover useful information and draw suggestive conclusions that help in decision-making processes related to network planning and operation.

During the project, two Data Analysis tools were used: Business Intelligence and Data Mining. The first is frequently described as a set of tools that transform raw data in meaningful and useful information. With the Data Mining tools, that information was explored so that consistent patterns could be identified.

As a result, a strong and well-founded methodology has been set up, making it possible to identify seven groups which characterize the behaviour of the MV network. For each group, the average load profiles by season and day type have been computed.
The analysis of the results showed a good adherence to the reality, taking into account the geographic location and the downstream loads of the PS that make up each group.

This work represents a starting point in the analysis of this kind of information, providing EDP Distribuição with a real knowledge of the network, essential for several business areas of the company, particularly in network planning, management and optimization.

This methodology of identification of electrical energy consumption or production patterns is being continued at EDP Distribuição for secondary substations, clients and producers.

The Typical Load Profiles, obtained with the work described, are the main input of probabilistic planning methods that are being implemented at EDP Distribuição.

REFERENCES

[1] P. Montero, J. Vilar, 2014, "TSclust: An R Package for Time Series Clustering", *Journal of Statistical Software*, vol. 62, 1-37.

[2] F. Iglesias, W. Kastner, 2013, "Analysis of Similarity Measures in Time Series Clustering for the Discovery of Building Energy Patterns", *Energies*, vol. 6, 579-597.

[3] Y. Zhao, 2013, *R and Data Mining: Examples and Case Studies*, Elsevier, 71-84.

[4] P. Manso, J. Vilar, 2015, *TSclust: Time Series Clustering Utilities*, R package version 1.2.3.

[5] U. Mari, A. Mendiburu, J. Lozano, 2015, *TSdist: Distance Measures for Time Series Data*, R package version 2.2.

[6] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, 1996, "A density-based algorithm for discovering clusters in large spatial databases with noise", *KDD*, 226-231.