Unsupervised classification of solar radiation daily data

Martín Gastón[1*] and Iñigo Pagola[1] and Lourdes Rámirez[1]

[1] National Renewable Energy Centre (CENER), Solar Thermal Energy Department, Ciudad de la Innovación 7, 31621 Sarriguren (Navarra), Spain

[*] Corresponding Author, mgaston@cener.com

Abstract

In this work we present the applicability of cluster analysis in the issue of determining the behaviour and structure of daily solar radiation in a specific emplacement.

In solar radiation assessment, classifications of the days according to the radiation integrate are often used. However it is known that this kind of classifications usually presents deficiencies. They don't collect all the aspects that could distinguish days, and also, they can classify days in the same group but presenting, for example, different behaviour in an energy production sense.

1. Introduction

Local meteorological conditions can provoke important variation of solar radiation and, therefore, these fluctuations must be taking into account in the previous studies of solar energy systems, particularly in solar thermal plants. Traditionally studies use daily or hourly data to study the radiation resource of an emplacement but don't use the respective information of these variations. The present work shows the first steps given with the goal of including the variability of the daily radiation in the resource studies. The idea is that, under the point of view of the plant operation, the daily typology of radiation is an important characteristic to know the future behaviour of the plant.

Then, the first action has been to study the inner structure present in the daily data. To do that, a cluster exercise has been made using the daily graphs of global irradiance as explanatory data. We show how a strong structure appears in our data and how the clusters can be easily interpreted.

The obtained results invite to research the mentioned inner structure of these kinds of data to determine the solar radiation behaviour of site thinking in the operation of a plant. The same integrate in a day can produce different energy production depending of the hourly distribution.

In the second section we make a brief introduction to the cluster analysis, after that we describe the data used in the present work and the obtained results. To finish we related some conclusions and future researches.

2. Cluster analysis

Cluster analysis covers all those algorithms whose aim is to group objects based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar to each other and different from the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the "better" or more distinct the clustering.

So, a cluster analysis can be divided into four steps:

• Data collection and selection of the variables for analysis: We must decide the variables which better describe our objects and therefore, those that would provide a structure in the population.

• Selection of a dissimilarity measure and generation of a similarity matrix: One of the keys in a cluster analysis is the use of a suitable dissimilarity measure. It is necessary a way to measure the proximity or the level of association between different objects. The similarity matrix collects all the differences between the available objects.

· Decision about number of clusters and interpretation: It depends on the algorithm, but in a lot of them, we must first select how many clusters compounds the data. After the construction of the clusters, it is usually necessary to make an interpretation of each group, that is, to assign a sense of each one of the built clusters.

• Validation of cluster solution: To finish, it is interesting to make a validation of the clusters. To do that, there are some statistics, as the silhouette, that describes the veracity of the found groups. Also, in some cases and depending on the nature of the data, it is possible to validate by means of a testing sample.

3. Application on solar radiation data

In this work we present the obtained results in a Spanish location sited in south of iberian peninsula. A year of hourly registers of global irradiance is available for the emplacement and after the typical quality tests we work with a set of 339 days.

The goal is to build groups of days whit similar features into each group and with large difference between groups. If the clustering exercise presents good properties under the point of view of the observed structure and an easy interpretation, then cluster could be seen as a start point in the site characterization task.

As we say before, the first task is to determine the variable that will describe our objects, in our case the objects of interest are de days under the point of view of radiation. In front of use the daily integrate we take, as explanatory variable of each day, the set of the hourly global irradiance registers. Then, we work with daily graphs of radiation as we can see in figure 1 when a sample of four data has been plotted.

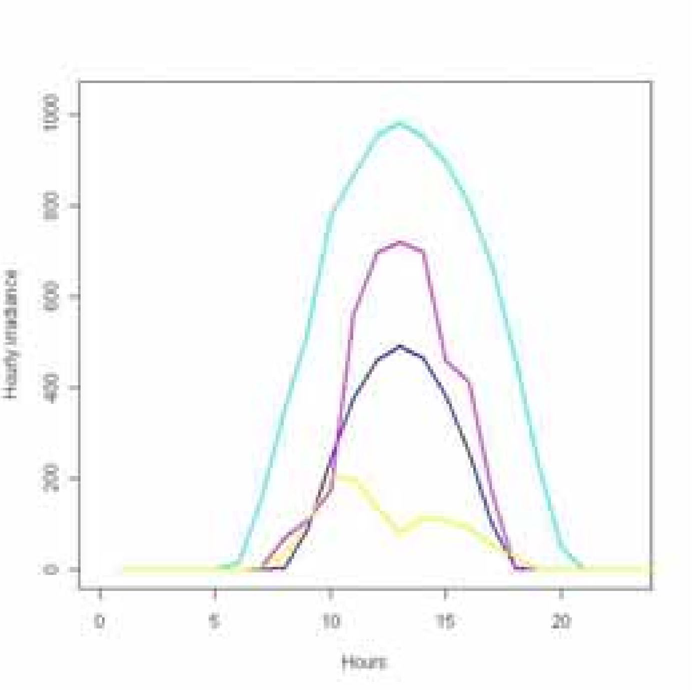

Fig. 1. Sample of data used to classify.

The idea is that taking into account all hourly registers of a day these functional data collects all daily behaviour and then, the cluster exercise will be more productive.

Given the nature of our variables, we must use appropriate tools to treat with them. So, we use methodologies born in the functional data framework ([1], [2] and [3]) to select the similarity measures between days. That is, the data of our study actually are functions and therefore we can use functional operators to make the clusters.

We check the suitability of different dissimilarity measures between functions as the L2-distance and the pca-distance with similar results. In this point we must note that it could be interesting the definition of a similarity measure ad-hoc to our problem and increase, for example, the robustness of the cluster structure. Also, design different measures can allow classify under different goals.

Besides, we use the PAM (partition around medoids) algorithm as cluster tool. The details of the algorithm can be seen in [4]; and after some attempts, as is described in [5], we decided to make the clusterization into three groups. To finish, as first criteria of validation we use the silhouette, a statistic that describes the suitability of the generated clusters whose mathematical expression can be seen in [6]. If the silhouette is greater than 0.4 the obtained classification collects a great structure. Other cases the cluster do not have a clear structure and it is possible that their interpretation were absurdity.

4. Results

In this section we expose the obtained results in the daily classification exercise. As we say in previous section, we decide that there are three clusters of days and by mean of the L2 distance and the PAM algorithm classify the data with an observed silhouette of 0.65, which indicates a greater structure in the clusters observed.

To help the cluster interpretation it is useful plotting a sample of each group and then make hypothesis about the sense of the cluster. Figure 2 collect one of these samples. We can see how data of the third group are summer clear days, in the second group we can see days with low radiation and the first group is formed by days with cloud pass and winter clear days.

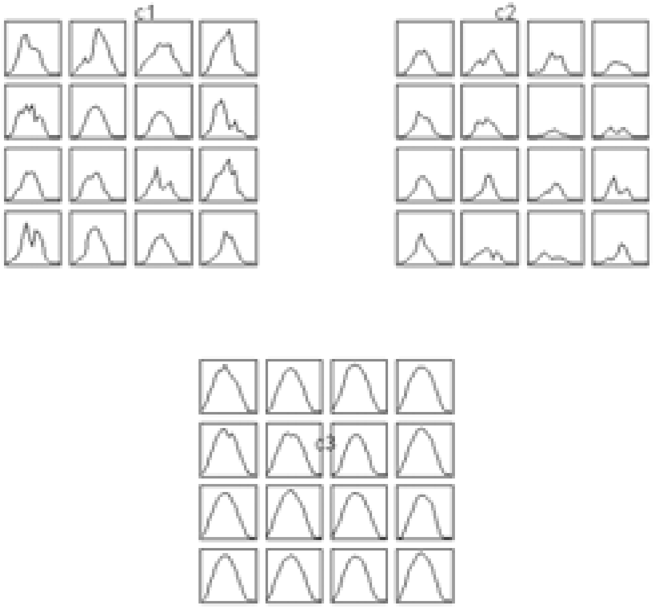

Fig. 2. Sample of data of each cluster.

5. Conclusions and future researches

After this first exercise of unsupervised classification of days we have observed that the data have a clear inner structure, acceptable in the sense that there are levels of silhouette greater that 0.65 and a good classification is accepted when this coefficient is greater than 0.4. Furthermore, the interpretation of the clusters is easy: one of them collects the clear days in summer, another one is formed by clear days in winter and slightly cloudy days, and the third one collects very cloudy days.

The job presented in this paper must be seen as a first approach to the daily classification task by mean of functional data analysis and unsupervised classification. The obtained results present these tools as a possible start point in the specific emplacement characterization, it is suggested the design of appropriate dissimilarities to increase the sense of the clusters. It also could be useful the use of greater frequency of registers, that is, to use ten minutes registers in front of hourly ones appears as a problem more suitable to be treating under the functional point of view.

References

[1] J.O. Ramsay, (2005). Functional Data Analysis, Springer, London.

[2] F.Ferraty, P. Vieu, (2008) Nonparametric Functional Data Analysis: Theory and Practice, Springer, London.

[3] G. Ayala, M. Gastón, T. León, F. Mallor (2008)Measuring Dissimilarity Between Curves by Means of Their Granulometric Size Distributions. Functional and Operatorial Statistics p. 35-41, Springer.

[4] M. J. Van Der Lann, K. S Pollard, J. E Bryan, (2003). "A New Partitioning Around Medoids Algorithm". *Journal of Statistical Computation and Simulation* (Taylor & Francis Group) **73** (8): 575–584.

[5] C. A. Sugar and G. M. James (2003). Finding the number of clusters in a data set: An information theoretic approach. Journal of the American Statistical Association 98 (January): 750–763.

[6] Peter J. Rousseeuw (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics 20: 53–65.

[7] J. A. Hartigan, (1975). Clustering Algorithms. Wiley.