Intelligent Techniques for Identification of Zones with Similar Wind Patterns

José Carlos Palomares-Salas[1,2], Agustín Agüera-Pérez[1,2],
and Juan José González de la Rosa[1,2]

[1] Research Group PAIDI-TIC-168: Computational Instrumentation and Industrial Electronics (ICEI)

[2] University of Cádiz. Area of Electronics. EPSA. Av. Ramón Puyol S/N.
E-11202-Algeciras-Cádiz-Spain

{josecarlos.palomares,agustin.aguera,juanjose.delarosa}@uca.es

Abstract. Two process to demarcate areas with analogous wind conditions have been developed in this analysis. The used techniques are based on clustering algorithms that will show us the wind directions relations for all stations placed in the studied zone. These relations will be used to build two matrixes, one for each method, allowing us working simultaneously with all relations. By permutation of elements on these matrixes it is possible to group related stations. These grouped distributions matrixes will be compared among themselves and with the wind directions correlation matrix to select the best algorithm of them.

Keywords: Cluster Analysis, Clustering Applications, Data Mining, Self-Organizing Feature Map.

1 Introduction

Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used, most of the times, in data mining, machine learning, pattern recognition, image analysis, bioinformatics or dimension reduction. However, because of so many problems, there is little prior information (e.g., statistical models) available related to the register of the data, and the decision-maker must do as few assumptions about the data as possible. It is under restrictions that the clustering methods are particularly appropriate for the exploration of the interrelationships among the data points to make an assessment (perhaps preliminary) of their structure [5]. In this analysis we propose its utilization to select areas with similar wind patterns.

The method used to compile and classify manually data is expensive, and the characterization of the patterns changes with time. On the other hand, it allows us to find useful characterization to build classifiers, and the discovery of class and subclass to reveal the nature of the structure problems.

There are many clustering techniques; the most worldwide used are hierarchical clustering and dynamic clustering [9]. The first one is called clustering tree and it is one of the most clustering approaches worldwide used, due to

M.L. Reyes et al. (Eds.): IT Revolutions 2011, LNICST 82, pp. 10–21, 2012.

the great visualization power it offers. Hierarchical clustering produces a nested hierarchy of similar groups of objects, according to a pairwise distance matrix of the objects. One of the advantages of this method is its versatility, because the operator does not need to provide any parameters such as the number of cluster. However, its application is limited to only small datasets, due to its quadratic computational complexity [6]. The second is the well known k-means. While the algorithm is perhaps the most commonly used clustering algorithm in the literature, it has several shortcomings, including the fact that the number of cluster must be specified in advance [2], [3]. Both of these approaching clustering, however, require that the length of each time series must be identical due to the Euclidean distance calculation requirement. They are unable to deal effectively with long time series due to its poor scalability. As in the supervised classification methods, there is not any clustering technique that is universally applicable.

The demarcation of different areas with connected wind patterns could have an important contribution to prediction models, based on data acquired in meteorological stations placed in the studied area. When these models are based on the statistical learning of data (Neural Networks, ARMAX, Genetic Fuzzy Learning), the inclusion of not correlated or erroneous stations can destabilize the process of getting the desired information.

In this article, we will use two clustering analysis techniques to classify zones with similar wind patterns. These techniques have been: the hierarchical clustering and the self-organizing feature map (SOFM). The main reasons of applying these algorithm are the capability of learning, for example held in SOFM model [7], and its great visualization power. One time that this first clustering has been done, we propose a new method based in Genetic Algorithms to optimize the final classification of the studied zone.

The paper is structured as follows. The following Section 2 presents the zone of study selected and the data used. Section 3 describes the clustering algorithms used in this paper. Section 4 is dedicated to the form of acquisition of the similarity matrices for these methods. Section 5 describes the optimization of the matrix of similarity with Genetic Algorithm. Finally, the results are presented in Section 6 and conclusions are drawn in Section 7.

2 Target Area and Wind Data

In this work the mean daily wind speed and direction of 88 met stations, from 2005 to 2008 have been used to select the reference stations. Data of the target station were acquired from a unit located in Southern of Andalusia (Peñaflor, Sevilla). The data from long-term wind for the target station contains measures taken at intervals of ten minutes and they cover the period since 2007 to 2008. The mean daily wind speed and direction were calculated from these data and are utilized for test procedure.

The map of the region and the location of these stations are depicted in Fig. 1. These stations are distributed over 87000 Km^2 and they are conceived to measure agriculture variables (Andalusia agriculture-climate information network).

In this way, wind records have not enough reliability because, despite of the most of them are located in open zones, the anemometer's height is $1.5m$ and is highly affected by obstacles and ground effects. (This fact add value to this study because this kind of meteorological records are more frequent than the good ones, and is interesting to build a structure that allows to use them in order to the wind resource evaluation.)

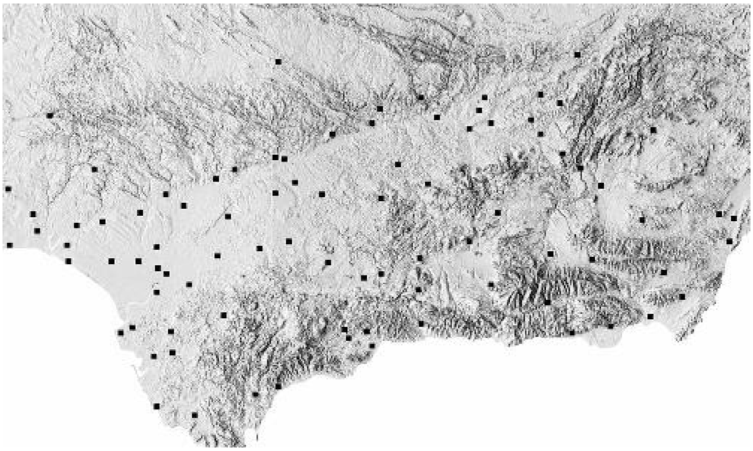

Fig. 1. Map of Andalusia showing the location of the meteorological stations

3 Methods for Exploratory Data Analysis

Several well-known methods are used for quickly producing and visualizing simple summaries of data sets. The main goal in the exploratory data analysis is to extract useful information out of large and usually high-dimensional data sets. Then, we present the characteristics of any methods proposed to classify high dimensional datasets. These methods have been: hierarchical clustering and Kohonen's self-organizing feature map ($SOFM$).

3.1 Hierarchical Clustering

The hierarchical clustering algorithms operate incrementally. Initially each data represent its own cluster. Then proceed successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The clustering methods differ in the rule by which it is decided which two small clusters are merged or which large cluster is split. The end result of the algorithm is a tree of clusters called a dendogram, which shows how the clusters are related. By cutting the dendrogram at a desired level a clustering of the data items into disjoint groups is obtained [4,1].

3.2 The Self-Organizing Map Algorithm

Kohonen's Self-Organizing Feature Maps represent a very well known and widely used Neural Network Model which employs unsupervised learning [8]. This means that the result is obtained from the properties inherent to the data itself; there is no a concise model that indicates whether it is true or false. So, no previous knowledge about the structure of the data is needed to be able to use the algorithm.

Self-organizing feature maps ($SOFM$) learn to classify input vectors according to how they are grouped in the input space. The neighbour neurons in the $SOFM$ learn to recognize neighbour sections of the input space. Therefore, the $SOFMs$ are trained to learn about the distribution (like in competitive layers) as well as the topology of the input vectors.

4 Proposed Procedure for Selecting Reference Stations Using the Different Techniques

Upon review some characteristics and properties of the used algorithms, we will put into practice them for to demarcate areas with similar wind patterns to target station within our study zone. For this purpose the following procedure has been executed:

4.1 Characterization of Reference Stations

The process followed to characterize the stations is function of used technique. The *Matlab R*2008*b* software has been our analytic tool.

Hierarchical Clustering Algorithm. The wind directions at all stations have been chosen for two random days yielding a vector of dimension (2xnS), where the number of stations (nS) is $nS = 89$. Left graph in Fig. 2 shows a dispersion graph where is show the pairwise of this vector.

Once we have obtained this vector, the hierarchical clustering algorithm has been carried out. This algorithm may be represented by dendrogram which illustrates the fusions or divisions made at each successive stage of analysis. For this propose we have been input the Euclidean distance as distance parameter to form the clusters and we apply as distance between clusters of 10. This algorithm will return a vector with the clusters associated to each pairwise of measurement. Right graph in Fig. 2 shows the connections of the first 20 clusters formed for that vector. This graph represent a snapshot of the relations among the stations reduced to the information of two random days. If two new days were chosen, the situation of the stations will change and the clusters will contain different elements. After n repetitions of the process, where n is equal to 90% of number of available data, we obtain a matrix of dimension (nxnS).

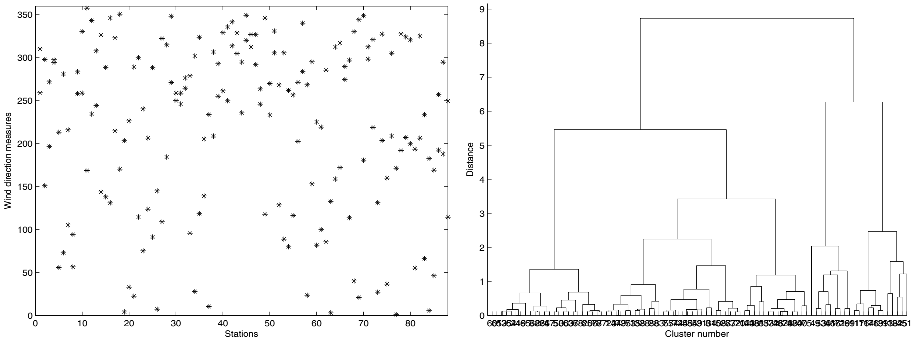

Fig. 2. Left: Dispersion Graph for measurements of wind directions for two random days in the zone of study. Right: Dendrogrma of a vector data.

After we obtain the matrix characteristic of the stations it is possible to determine how many times two stations have been inserted in the same cluster. The higher the number of coincidences, the more similarity between the wind patterns in both locations. Where n_ij is the parameter that represents the number of coincidences of the i-station and the j-station, it is proposed Eq. 1 defined in the range $[0, 1]$ as a measure of the similarity between their wind patterns.

$$S_{ij} = \frac{n_{ij}}{n} \tag{1}$$

Calculating this parameter for all possible pairs of stations, the matrix S (composed of S_{ij}) can be constructed with dimension (nSxnS).

***SOFM* Algorithm.** To characterize the measurement stations are chosen at random wind directions (WD) of all them for two days to yield a vector of dimension 2xnS. Left graph in Fig. 3 depicts a graph of dispersion where is show the pairwise of this vector. The $y - axis$ represents the wind directions in degrees of all station for one random day and $x - axis$ represents the wind directions in degrees of all station for another random day.

Once we have obtained this vector, the SOFM algorithm has been carried out. The configuration parameters related to the implementation of SOFM by *Matlab R*2008*b* software were as follows: size of layer dimension, [2x5]; network topology, '*hextop*'; distance function, '*linkdist*'. With these parameters we perform the training of the network to obtain the weights of the network. Right graph in Fig. 3 shows the scatter plot together with the network's initial weights, and its corresponding finals weights after training.

Once the network has been trained we come back to choose two random days and is simulated with the obtained network configuration. Then we take note of what cluster it belongs each station. This process is repeated 373 times with what a matrix of size ($89x373$) is obtained to end. After this matrix is obtained, it is possible to determine how many times two stations have been inserted en the same cluster.

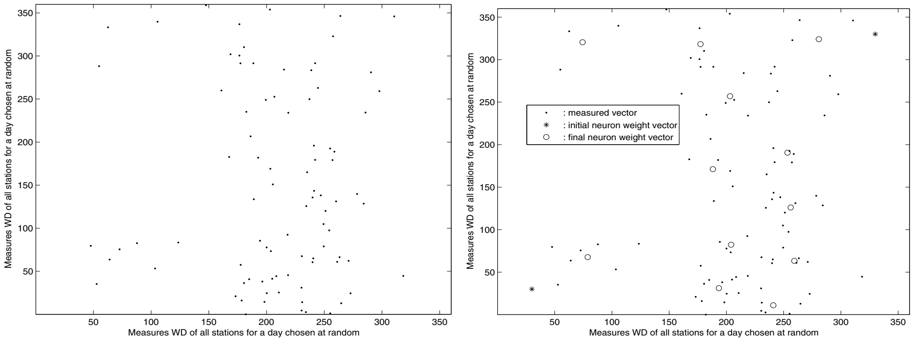

Fig. 3. Left: Dispersion Graph for measurements of wind directions for two random days in the zone of study. Right: SOFM training result where we maintain the outliers to remark that dispersion is low.

When all the stations are characterized by the position of the network weights, it is possible to established an indicate parameter degree of similitude (S) between two distributions. We propose the calculation of the mean distance between each neuron (Eq. 2), C, and the nearest one belonging to the other station distribution, C'_i. Thus, the lowest is value S; and the highest is the similarity of both distributions. The resulting matrix has dimension (nSxnS).

$$S_{ij} = \frac{\sum_n |C_n^i - C_n^j|}{n} \tag{2}$$

Correlation Matrix. The wind directions at all stations are chosen to form a matrix. Then, we calculate the pairwise linear correlation coefficient between each pair of columns from this matrix, obtaining at the end a correlation matrix with dimension (nSxnS). This matrix will help us to compare the used techniques later in the section 6.

5 Ordering the Matrixes *S* with Genetic Algorithm

The obtained matrixes in the previous section contain the relations among all the wind patterns measured at the stations, and it can be represented grouped by provinces. Fig. 4 shows the corresponding matrix to obtained in *SOFM* algorithm. The order of grouping of the provinces has been the following: Almería (Alm), Cádiz (Cad), Córdoba (Cor), Granada (Gra), Huelva (Hue), Jaen (Jae), Málaga (Mal), and Sevilla (Sev). The dark pixels are associated to a low value of S; therefore, they connect stations with similar patterns. Thus, the white cross observed over Málaga (Mal) stations indicates that the most of them have not relations with other stations, even if they are placed in the same province. On the contrary, Huelva (Hue) shows strong relations among the stations installed in the area. Córdoba (Cor) presents the same pattern in almost all the province, but this pattern is repeated in Sevilla (Sev), as it is possible to infer from the dark areas connecting these provinces. This fact indicates that the classification of the stations according to their provinces is not the best in order to visualize the areas with a similar wind patterns.

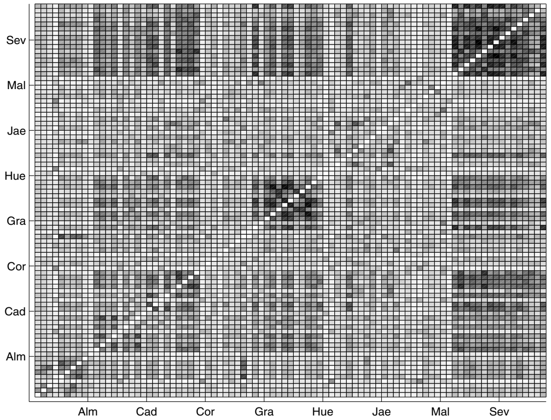

Fig. 4. Representation, in grey scale, of the matrix composed of the values of S for each pair of stations. Each station is labeled as an Andalusia province.

The actual order of the matrixes comes from alphabetical and administrative criteria, but these considerations have not relation with the concerned problem, the wind classification. If the stations were grouped according to the relations among them, by permutation of rows and columns of the matrix, the relations and clusters could be clarified.

Although the permutation of rows and columns to put in order the S matrixes seems to be a simple problem; the reality proves that this process could be compared with a Rubik cube, since the order in a part of the matrixes could involve the disorder in other one.

The result (or objective) of the recombination of rows and columns must be a matrix in which the stations with similar winds patterns and relations will be neighbors, that is, the nearby elements of the obtained matrix must be as similar as possible. Fig. 5(left) and Fig. 5(right) present two possible recombination of the matrix represented in Fig. 4, being the second one closer to the objective explained before. To evaluate this idea of order, the parameter p is proposed in Eq. 3, where p_0, a and b are constants related to the scale of the problem. In this case $p_0 = 25000$, $a = 100$ and $b = 415$.

$$p = \frac{1}{p_0} \cdot \sum_{j=1}^{88} \sum_{k=-3}^{k=3} \sum_{i=1}^{88} F_{ij} \cdot (A_{ijk} + B_{ijk}) \tag{3}$$

$$F_{ij} = 1 - \frac{|i-j|}{88} \tag{4}$$

$$A_{ijk} = \frac{a}{a + (S_{ij} + S_{i(j+k)})} \tag{5}$$

$$B_{ijk} = \frac{|S_{ij} - S_{i(j+k)}|}{b} \tag{6}$$

The $j - th$ column, which represents a station, is compared with the six closer columns indexed by $j + k = j - 3, ..., j, ...j + 3$, calculating two factors with their $i - th$ elements, A_{ijk} and B_{ijk}. The resulting value of $A_{ijk} + B_{ijk}$ is low when the sum of the elements is high and the difference low. That is, nearby stations with high similarities among them and with analogous relations with the rest of the stations will contribute with low values to the final result of p. The sum of all these values, covering all the columns, gives an objective measurement of the similarities among the nearby columns and, therefore, an evaluation of the global order of the matrix. For example the value of p for the matrix shown in Fig. 4 is 1.430. As it was expected, Fig. 5(left) and Fig. 5(right) obtain lower values because they have been ordered in some sense. Especially the combination represented in Fig. 5(right) presents a very low value of p ($p = 0.844$) which indicates a high degree of similarity (or order).

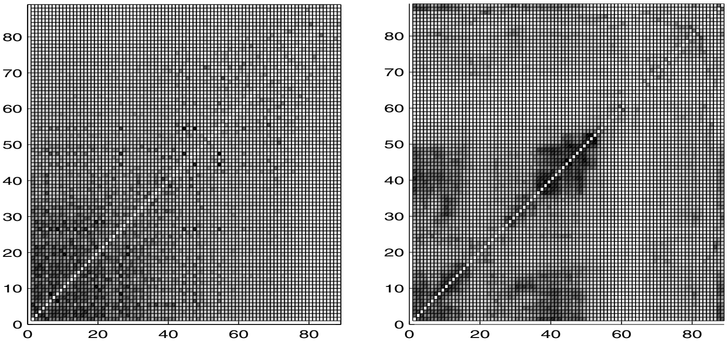

Fig. 5. Left: Ordination of stations from sparsely related to highly related ($p = 0.989$). Right: Ordination according to subjective criteria of permutation ($p = 0.844$).

Now the problem of ordering the matrixes of similarities has been reduced to find a combination of stations with a minimum value of p. We propose to solve this minimization problem using Genetic Algorithms (GA). Each matrix of similarities can be characterized by a vector of 89 elements containing the position of the stations. This vector could be considered as a genome which defines univocally the associated matrix. Furthermore, using the value of p calculated with these matrixes, a population of these vectors could be tested and ranked. Under these conditions, GA could improve this population using evolutive operators as crossover, mutation, migration, etc., in order to obtain the minimum value of p.

As it has been introduced upper, the vectors used as genome of the matrixes contain 89 elements. These elements are non repeated integer numbers between 1 and 89, and each of them is associated to one of the used stations. The positions of these numbers in the vector define the position of the stations in the matrix and, thus, the value of p for this combination can be calculated. Because of the properties of the genome used in this work, the evolutive operator selected to produce the new generations is the Recombination. Recombination permutes one or more elements of the genome, thus, the resulting vector is composed of 89 non repeated integers again; avoiding the repetitions, decimals and values out of range given by other operators.

6 Results

Once the matrixes have been selected by the GA as best combination of stations, after 1000 generations and a population of 10^5 individuals, these are represented in Figs. 6(left) and 6(right), where have been selected the major clusters with numbered square. The same representation for the correlation matrix is showing in Fig. 7.

Table 1 shows the information of the stations that have been selected as cluster belong in the two used techniques, and the table 2 presents the selected stations in the correlation matrix.

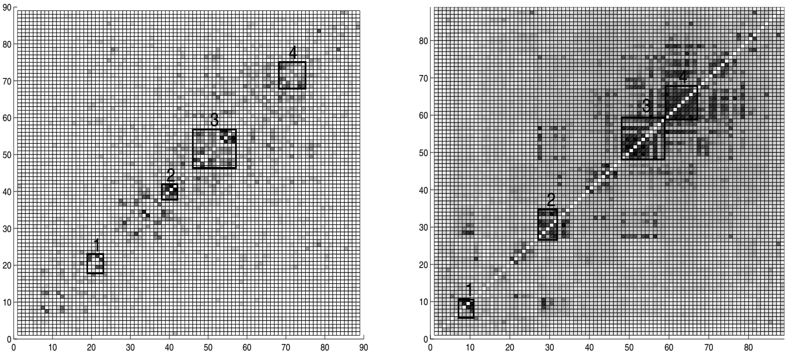

Fig. 6. Left: Representation, greyscale, of the matrix S belonging to hierarchical algorithm after applying the genetic algorithm. Right: Representation, greyscale, of the matrix S belonging to SOFM algorithm after applying the genetic algorithm.

The main objective in this study is to obtain reference stations with similar wind patterns to target station. The membership cluster of target station in table 2 is formed of 41 stations, but this cluster is too large. By contrast, the membership cluster of target station in the two used techniques is the number 3 and, besides, these clusters overlap by about 82% as we can see in table 1.

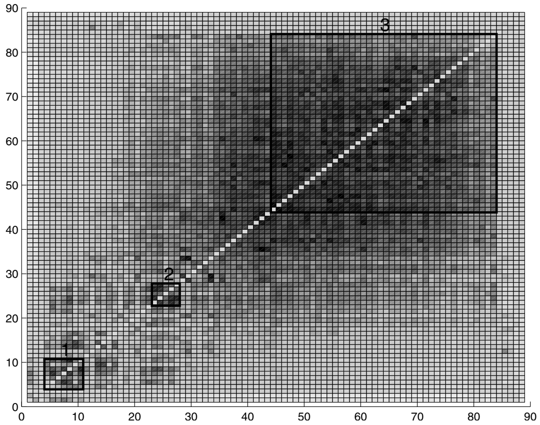

Fig. 7. Representation, greyscale, of the matrix S belonging to correlation algorithm after applying the genetic algorithm

Table 1. Names of the stations selected in the two clustering techniques used

Cluster	Hierarchical Clustering	SOFM Clustering
1	Torreblascopedro, Sierra Yeguas, Jaén, Mancha Real	Sanlúcar de Barrameda, La Puebla de Guzmán, Lepe, Gibraleón
2	Lepe, El tojalillo-Gibraleón, Ifapa el Cebollar, Moguer	Almonte, Ifapa el Cebollar, El Campillo, El tojalillo-Gibraleón, La Palma del Condado
3	Guillena, Almonte, Sanlúcar la Mayor, Linares, Aznalcázar, La Rinconada, Peñaflor, Lebrija, La Puebla del Río II, La Puebla del Río, Isla Mayor	Guillena, Isla Mayor, La Puebla del Río, La Puebla del Río II, La Rinconada, Lebrija, Aznalcázar, Hornachuelos, Sanlúcar la Mayor, Peñaflor, Palma del Río
4	La Luisiana, Los Molares, Las Cabezas de San Juan, Córdoba, El Carpio, Écija, Puebla Cazalla, Lora del Río	La Luisiana, Lora del Río, Puebla Cazalla, Villanueva del Río y Minas, Santaella, Torreblascopedro, Écija, Osuna

These techniques also have an overlap 50% in the clusters 2 and 4, we can say that these methods are able to identify areas with similar wind patterns more accurately than correlation matrix.

Between of two seen methods, the one based in *SOFM* algorithm is more appropriate to determine areas with similar wind patterns because this is a more robust statistical classification method than the other one.

Table 2. Names of the stations selected applying the correlation matrix

Cluster	Correlation
1	Cuevas de Almanzora, Cártama, Estepota, Churriana, Fiñana, Almería, Tíjola
2	Puebla de Don Fabrique, Huesa, Cádiar, Pozo Alcón, San José de los Propios
3	Villanueva del Río y Minas, Lora del Río, Mancha Real, Hornachuelos, Santaella, Córdoba, La Palma del Condado, El Carpio, Sierra Yeguas, Sanlúcar la Mayor, Torreblascopedro, La Luisiana, Linares, Puebla Cazalla, Peñaflor, La Rinconada, Moguer, Palma del Río, Guillena, Isla Mayor, La Puebla del Río II, Finca Tomejil, Écija, Aznalcázar, La Puebla del Río, Puerto Santa María, Adamuz, Lebrija, Osuna, Los Molares, Las Cabezas de San Juan, Baena, Basurta-Jerez, Ifapa Centro Cabra, Alcaudete, Santo Tomé, Jaén, Sabiote, La Higuera de Arjona, Ubeda

7 Conclusions

In this study the used reference stations are low-quality stations and wind records are not reliable enough because although most of them are located in open areas, the height of the anemometer is 1.5 m and it is greatly affected by obstacles and ground effects. The target station measurements are taken at a height of 20 m and for this cause its have little correlation with the reference stations in wind speed.

The results obtained in the first phase show that the proposed method is able to demarcate areas with analogous wind patterns, even if the data acquired is affected by low quality instruments or locations. In the same way, erroneous stations, or stations not representative of the wind climate in their zone, will be identified since they will not be included in any cluster. So, this tool could be useful in two aspects:

- In first steps of wind resource assessment, when a preliminary description of the wind climate in a zone is needed. Then, using the information given by this matrix, it is possible to associate the location of the target area with an expected wind pattern.
- When a wind methodology, as Measure-Correlate-Predict or the ones used in wind temporal forecasting, needs support stations to complete or extend the database used. In this situation is very important to exclude stations with errors or not representative of the studied area because it could lead to important differences between results and reality.

It can be concluded that data from low-quality stations can help to make predictions in a target station. First we would have to carry out a phase of treatment of low-quality data to consolidate the database eliminating gaps or using any technique for filling. This can help us execute the study of wind potential in the area, being able to reduce the costs associated to data acquisition phase due to the large amount of low-quality data available.

Acknowledgement. The authors would like to thank the *Spanish Ministry of Science and Innovation* for funding the research projects *TEC*2009-08988 and *TEC*2010-19242-*C*03-03 (*SIDER-HOSAPQ*), and the European Program *FEDER-FSE* 2007 − 2013 for funding the teaching and research staff in universities. Our unforgettable thanks to the trust we have from the *Andalusian Government* for funding the Research Group $PAIDI - TIC - 168$ in *Computational Instrumentation and Industrial Electronics-ICEI.*

References

1. Anderberg, M.R.: Cluster analysis for applications. Probability and Mathematical Statistics (1973)
2. Bradley, P.S., Fayyad, U.M.: Refining initial points for k-means clustering. In: Proceedings 15th International Conference on Machine Learning, Madison, USA, pp. 91–99 (1998)
3. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Intelligent Information Systems 17, 107–145 (2001)
4. Hartigan, J.A.: Clustering algorithms. Journal of Classification (1975)
5. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys 31, 264–323 (1999)
6. Keogh, E., Lin, J., Truppel, W.: Clustering of time series subsequences is meaningless: Implications for past and future research. In: Proceedings 3rd IEEE International Conference on Data Mining, Melbourne, USA, pp. 115–122 (2003)
7. Kun-Lin, H., Cheng-Chang, J., I-Ching, Y., Yan-Kwang, C., Chun-Nan, L.: The study of applying a systematic procedure based on sofm clustering technique into organism clustering. Expert Systems with Applications: An International Journal 33, 330–336 (2007)
8. Lippmann, R.: An introduction to computing with neural nets. IEEE ASSP Magazine 4, 4–22 (1987)
9. Wang, X., Smith, K., Hyndman, R.: Characteristic-based clustering for time series data. Data Mining and Knowledge Discovery 13, 335–364 (2006)