European Association for the Development of Renewable Energies, Environment and Power Quality (EA4EPQ)

International Conference on Renewable Energies and Power Quality
(ICREPQ'10)
Granada (Spain), 23th to 25th March, 2010

A new method for identification of zones with similar wind patterns using Hierarchical Clustering Techniques

J. C. Palomares Salas, A. Agüera Pérez, J. J. G. de la Rosa, J. G. Ramiro
Research Unit PAIDI-TIC-168. University of Cadiz. Electronic Area. Escuela Politécnica Superior
Avda. Ramón Pujol, S/N. E-11202-Algeciras-Cádiz (Spain)
Phone/Fax number:+0034 956 028020, e-mail: josecarlos.palomares@uca.es, agustin.aguera@uca.es

Abstract. In this paper it is shown a process to demarcate areas with analogous wind conditions. For this purpose a dispersion graph between wind directions will be traced for all stations placed in the studied zone. These distributions will be compared among themselves using the hierarchical clustering algorithm. This information will be used to build a matrix, letting us work with all relations simultaneously. By permutation of elements in this matrix it is possible to group relationed stations.

Keywords

Cluster Analysis, Clustering Applications, Data Mining, Unsupervised Learning, Machine Learning.

1. Introduction

Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used, the majority of times, in data mining, machine learning, pattern recognition, image analysis, bioinformatics or dimension reduction [1]. However, in many such problems, there is a little prior information (e.g., statistical models) available about the data, and the decision-maker must make as few assumptions about the data as possible. It is under restrictions that clustering method is particularly appropriate for the exploration of interrelationships among the data points to make an assessment (perhaps preliminary) of their structure [2]. Here we propose its utilization for to selection regions where similar conditions exist.

This method is used when to compile and classify by hand is expensive, and the characterization of the patterns change with time. On the other hand, lets to find useful characterization to build classifiers, and the discovery of class and subclass that to reveal the nature of the problem structure.

There are many clustering techniques; the most widely used are hierarchical clustering and dynamic clustering [3]. The first are the called clustering tree and is one of the most widely used clustering approaches due to 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 generality, since the user 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 [4]. The second is the well knows *k*-means. While the algorithm is perhaps the most commonly used clustering algorithm in the literature, it does have several shortcomings, including the fact that the number of cluster must be specified in advance [5], [6]. Both of these clustering approaches, however, require that the length of each time series be identical due to the Euclidean distance calculation requirement, and are unable to deal effectively with long time series due to poor scalability. As in the supervised classification methods, there is not clustering technique that is universally applicable.

The demarcation of different zones 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 knowledge.

The remainder of the paper is organized as follows. The 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 management of the matrix of similarity with Genetic Algorithm. Finally, the section 6 concludes the paper and outlines some directions for future research.

2. Area and Wind Data

In this work the daily mean wind speed and direction of 88 met stations, from 2005 to 2008 have been used. These stations are distributed over 87000 Km^2 and they are orientated to measure agriculture variables (*Red de Información Agroclimática*). In this way, wind records have not enough reliability because, despite of the most of them are located in open zones, the anemometer height is 1,5 m 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.)

3. Hierarchical Clustering Algorithm

Cluster Analysis, also called data segmentation, has a variety of goals. All relate to grouping or segmenting a collection of objects (also called observations, individuals, cases, or data rows) into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters. Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered. There are two major methods of clustering: hierarchical clustering and k-means clustering.

In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to *n* clusters each, containing a single object. Hierarchical Clustering is subdivided into *agglomerative* methods or "bottom up", which begins with each element as a separate cluster and merges them into successively larger clusters, and *divisive* methods or "top down", which begin with the whole, set and proceed to divide it into successively smaller clusters.

A.1. Agglomerative method

An agglomerative hierarchical clustering procedure produces a series of partitions of the data, P_n, P_{n-1},…, P_1. The first P_n consists of *n* single object 'clusters', the last P_1, consists of single group containing all *n* cases.

At each particular stage the method joins together the two clusters which are closest together (most similar). (At the first stage, of course, this amounts to joining together the two objects that are closest together, since at the initial stage each cluster has one object.)

Differences between methods arise because of the different ways of defining distance (or similarity) between clusters. Usually the distance between two clusters *A* and *B* is one of the following:

Complete linkage clustering: Distance between groups is now defined as the distance between the most distant pair of objects, one from each group [7]. In this method, ***D(A,B)*** is computed as:

$$\max\{d(x,y) : x \in A, y \in B\} \tag{1}$$

Here the distance between every possible object pair (*x*,*y*) is computed, where object *x* is in cluster *A* and object *y* is in cluster *B* and the maximum value of these distances is said to be the distance between cluster *A* and *B*. In other words, the distance between two clusters is given by the value of the longest link between the clusters.

At each stage of hierarchical clustering, the clusters *A* and *B*, for which ***D(A,B)*** is minimum, are merged.

Single linkage clustering: The defining feature of the method is that distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered [8]. In this method ***D(A,B)*** is computed as:

$$\min\{d(x,y) : x \in A, y \in B\} \tag{2}$$

The distance between two clusters is given by the value of the shortest link between the clusters.

Average linkage clustering: Here the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group [9]. The distance ***D(A,B)*** is computed as:

$$\frac{1}{|A|\cdot|B|}\sum_{x\in A}\sum_{y\in B} d(x,y) \tag{3}$$

Average group linkage: With this method, groups once formed are represented by their mean values for each variable, that is, their mean vector, and inter-group distance is now defined in terms of distance between two such mean vectors [10]. In the average group linkage method, the two clusters *A* and *B* are merged such that, after merger, the average pairwise distance within the newly formed cluster, is minimum. Suppose we label the new cluster formed by merging clusters *A* and *B*, as *C*. Then **D(*A*,*B*)**, the distance between clusters *A* and *B* is computed as:

$$average\{d(x,y) : x, y \in C\} \tag{4}$$

At each stage of hierarchical clustering, the clusters ***A*** and ***B***, for which **D(A,B)** is minimum, are merged. In this case, those two clusters are merged such that the newly formed cluster, on average, will have minimum pairwise distances between the points in it.

Ward´s hierarchical clustering method: Ward proposed a clustering procedure seeking to form the partitions P_n, P_{n-1},…, P_1 in a manner that minimizes the loss associated with each grouping, and to quantify that loss in a form that is readily interpretable [11]. At each step in the analysis, the union of every possible cluster pair is considered and the two clusters whose fusion results in minimum increase in 'information loss' are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion, ESS.

Each agglomeration occurs at a greater distance between clusters than the previous agglomeration, and one can decide to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion).

Upon review some methods of hierarchical clustering algorithm we will implement it to demarcate areas within our study zone. To perform this type of analysis in our problem the following procedure has been executed: We choose the wind directions at all stations for two random days yielding a vector of dimension 2x88. The figure 1 depicts a graph of dispersion where is show the pairwise of this vector.

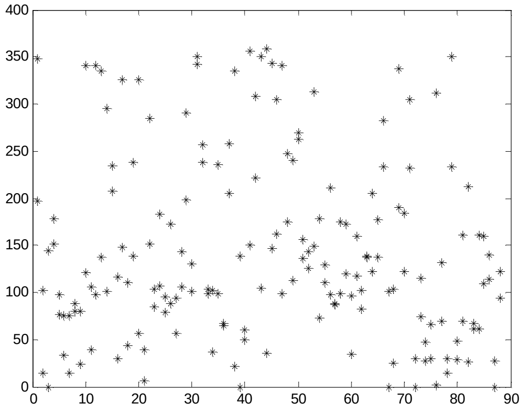

Fig.1. Graph of dispersion for measurement of two random days in the zone of study.

Once we have obtained this vector we apply hierarchical clustering algorithm. This algorithm may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis. For this propose we has been input the Euclidean distance as distance parameter for 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 pair of measurement which is shown in the figure 2. The figure 3 shows the connections of the first 20 clusters formed for that vector.

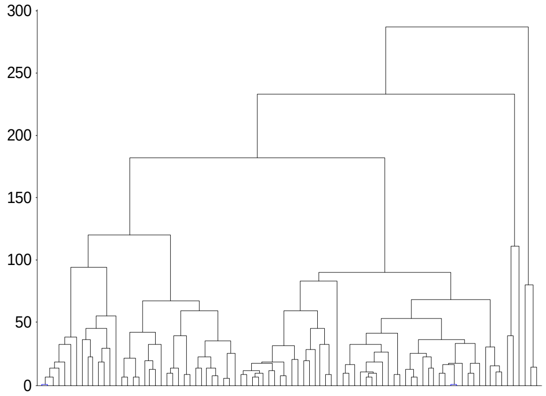

Fig.2. Dendrogram for vector data shown in figure 1.

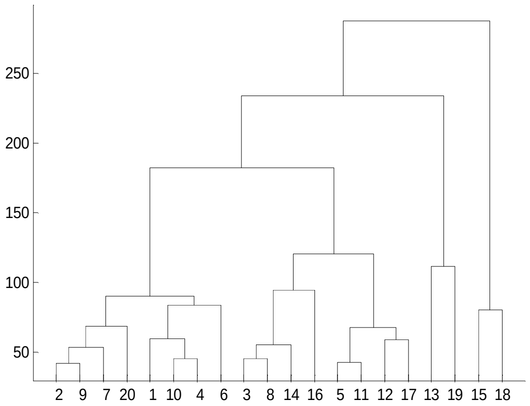

Fig.3. Detail of Dendrogram indicated in figure 2 showing the connections of the first 20 clusters.

4. Matrix of similarities

The figure 2, showed in the previous section, represents 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, 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. Let n_{ij} represent the number of coincidences of the *i*-station and the *j*-station. We propose S_{ij} (5), defined in the range [0, 1], as a measurement of the similarity between their wind patterns:

$$S_{ij} = \frac{n_{ij}}{n} \tag{5}$$

Calculating this parameter for all possible pairs of stations, the matrix S (composed of S_{ij}) can be constructed. This matrix contains the relations among all the wind patterns measured at the stations, and it can be represented as figure 4 shows, grouped by provinces. The order of grouping of the provinces is *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.

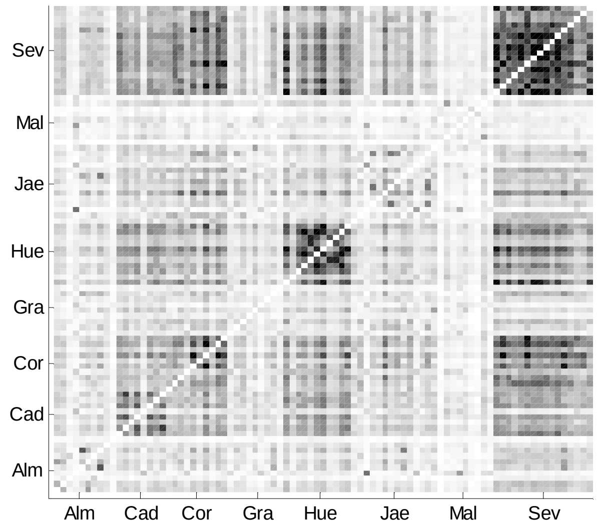

Fig.4. Representation, in grey scale, of the matrix composed of the values of S for each pair of stations.

The actual order of the matrix 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.

5. Ordering the matrix S with Genetic Algorithm

Although the permutation of rows and columns to put in order the S matrix 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 matrix 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 neighbours, that is, the nearby elements of the obtained matrix must be as similar as possible. Figures 5a and 5b present two possible recombinations of the matrix represented in figure 4, being the second one closer to the objective explained before. To evaluate this idea of order, the parameter *p* is proposed in equation 6, where p_0, *a* and *b* are constants related to the scale of the problem. In this case $p_0 = 25000$, $a = 100$ and $b = 415$.

$$p = \frac{1}{p_0} \cdot \sum_{j=1}^{88} \sum_{k=-3}^{k=3} \sum_{i=1}^{88} F_{ij} \cdot \left(A_{ijk} + B_{ijk}\right) \qquad (6)$$

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

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

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

Each column, *j*, 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 figure 4 is 0.902. As it was expected, figures 5a and 5b obtain lower values because they have been ordered in some sense. Especially the combination represented in 5b presents a very low value of *p* (*p*=0.843) which indicates a high degree of similarity (or order).

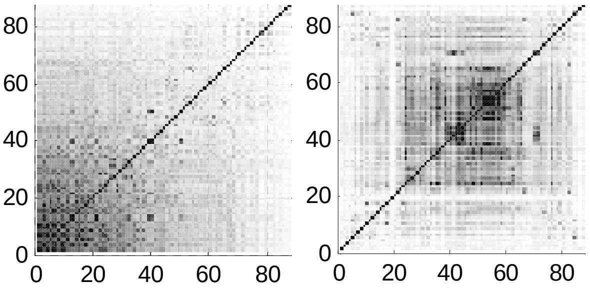

Fig. 5a) Ordination of stations from sparsely relationed to highly relationed (*p*=0.854). 5b) Ordination according to subjective criteria of permutation (*p*=0.843).

Now the problem of ordering the matrix 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 88 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. In 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 88 elements. These elements are non repeated integer numbers between 1 and 88, and each of them is associated to one of the used stations. The positions of this 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 88 non repeated integers again; avoiding the repetitions, decimals and values out of range given by other operators (Figure 6).

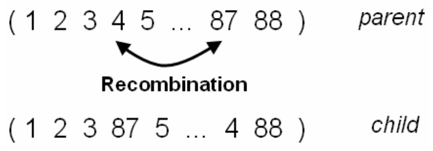

Fig.6. Recombination of one permutation.

5. Results

The matrix selected by the GA as best combination of stations, after 1000 generations and a population of 10^5 individuals is represented in figure 7. The value of *p* associated to this matrix is 0.805.

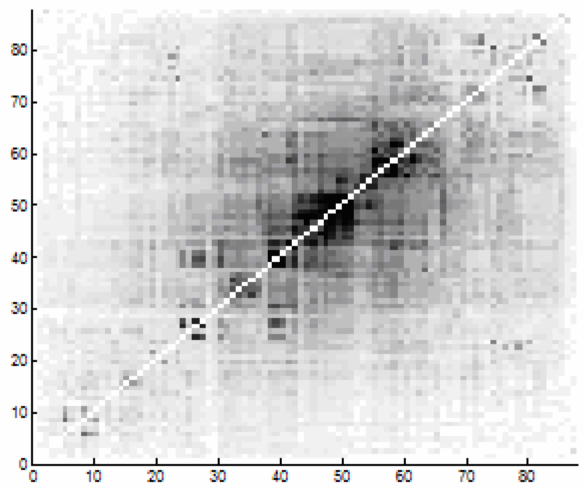

Fig.7. Representation, in grey scale, of the matrix obtained before to apply the Genetic Algorithm.

Figures 8 is selected the major clusters with colours, and figure 9 shows the same, but where the cluster have more definition.

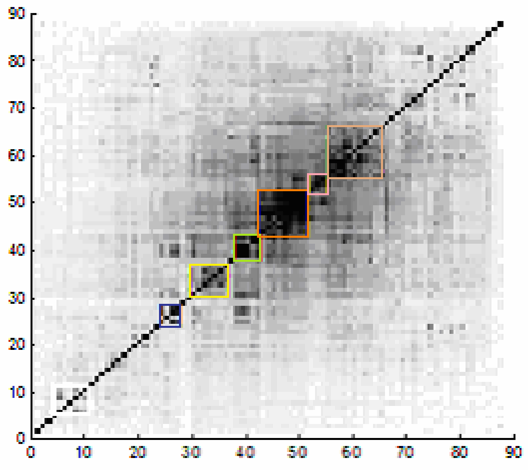

Fig.8. Illustration, with colours, of major clusters.

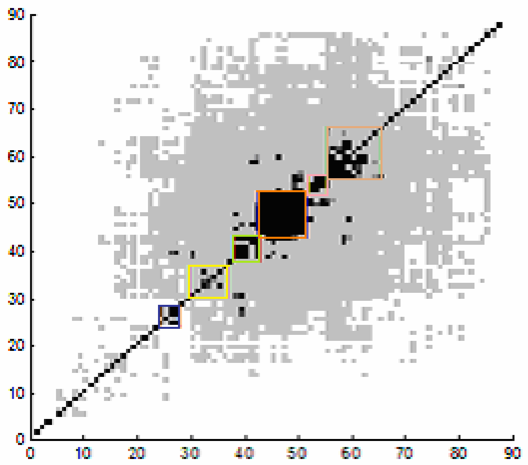

Fig.9. Image of the major clusters but where the clusters are represent with more definition.

Once we have selected the clusters are represented in the study area. This is shown in figure 10 where it follows the same colour code. Table I shows the information of the stations that have been selected as cluster belong.

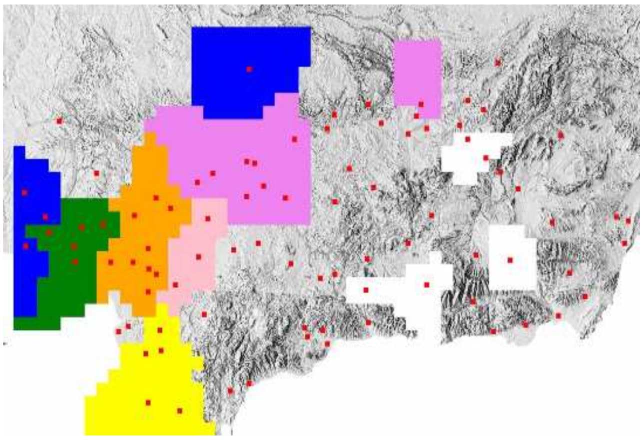

Fig.10. Representation of study zone and the clusters selected.

Table I. – Names of the stations selected in the clusters.

La puebla de Guzmán	1	Blue
Bélmez	1	
Gibraleón	1	
Lepe	1	
Conil de la Frontera	2	Yellow
Jerez de la Frontera	2	
Puerto Sta Mª	2	
Vejer de la Frontera	2	
Basurta-Jerez	2	
Moguer	3	Green
Ifapa El Cebollar	3	
El Tojalillo-Gibraleón	3	
Niebla	3	
La Palma del Condado	3	
Sanlúcar la Mayor	4	Orange
Guillena	4	
Almonte	4	
La Rinconada	4	
Aznalcázar	4	
La Puebla del Rio	4	
Lebrija	4	
La Puebla del Río II	4	
Isla Mayor	4	

La Luisiana	5	Magenta
Palma del Rio	5	
Écija	5	
Hornachuelos	5	
Lora del Rio	5	
Santaella	5	
Cordoba	5	
Villanueva del Rio y Minas	5	
Linares	5	
Finca Tomejil	6	Pink
Los Molares	6	
Las Cabezas de San Juan	6	
Huesa	7	White
Padul	7	
Zafarraya	7	
Fiñana	7	
San José de los Propios	7	

6. Conclusion

The results obtained demonstrate 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.

References

[1] Leif E. Peterson, Matthew A. Coleman, “Comparison of Gene Identification Based on Artificial Neural Network Pre-processing with k-Means Cluster and Principal Component Analysis”, Lecture Notes in Computer Science (2006), Vol. 3849, pp. 267–276.

[2] Jain, A.K., Murty, M.N., and Flynn, P.J. “Data clustering: A review”. ACM Computing Surveys (1999), 31(3): 265–323.

[3] Xiaozhe Wang, Kate Smith, Rob Hyndman. “Characteristic-Based Clustering for Time Series Data”, Data Mining and Knowledge Discovery (2006), Vol. 13, number 3, pp. 335–364.

[4] Keogh, E., Lin, J., and Truppel, W. “Clustering of time series subsequences is meaningless: Implications for past and future research”. In Proc. of the 3rd IEEE International Conference on Data Mining, Melbourne, FL, USA (2003), pp. 115–122.

[5] Bradley, P.S. and Fayyad, U.M. “Refining initial points for k-means clustering”. In Proc. of the 15th International Conference on Machine Learning, Madison, WI, USA (1998), pp. 91–99.

[6] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. “On clustering validation techniques”. Journal of Intelligent Information Systems (2001), 17(2/3): 107–145.

[7] Dawyndt P, De Meyer H, De Baets B. “The complete-linkage clustering algorithm revisited”. Soft Comput (2006) 9:385–392.

[8] Hartigan J. “Clustering algorithms”. Wiley, New Cork (1975).

[9] Everitt B. “Cluster analysis”. Wiley, New York (1993).

[10] Johnson, S.C. “Hierarchical Clustering Schemes”. Psychometrika (1967), 32: 241-254.

[11] Ward, J. H. JR. “Hierarchical grouping to optimize an objective function”. J. Am. Stat. Assoc. (1963). 58, 236–244.