CLUSTER ANALYSIS IN SEARCH OF WIND IMPACTS ON EVAPORATION

V. GÜLDAL[1] — H. TONGAL[2]

[1]*S.D.U. Faculty of Engineering and Architecture, Civil Engineering Department, Isparta, Turkey*

[2] *S.D.U. Natural and Applied Sciences, Civil Engineering Department, Isparta, Turkey*

**Corresponding author*
e-mail: vguldal@mmf.sdu.edu.tr

(Received 22nd February 2008; accepted 12nd June 2008)

Abstract. Clustering deals with finding a structure in a collection of uncategorized data and can be examined the most important unsupervised learning problem and the other problems as kind of this. The aim of this study is to cluster the monthly evaporation losses with the monthly winds speed and wind blow number of Eğirdir Lake, one of the most important fresh water storage of Turkey. For this aim, wind speed and evaporation data also wind blow number depend on hourly and daily mean records measured in Eğirdir Lake Catchment, are used. In the clustering analysis of the data, as a non-parametric approach hierarchical clustering algorithm was successfully applied at different similarity stages.
Keywords*: Clustering analysis,* h*ierarchical clustering algorithm, Eğirdir Lake, wind speeds and blow number, evaporation.*

Introduction

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects, similar between themselves and dissimilar to objects of other groups. By representing data with fewer groups, indispensable certain details can be lost but a simplification state can be achieved. Clustering can be considered as an unsupervised method for classification. If there is no prior information on the labels of the data (i.e. in which class they are), clustering algorithms determine the data to a usually pre-specified number of clusters (each cluster represented by a different stage). Clustering algorithms have been applied to a wide range of topics and areas. Uses of clustering techniques can be found in statistics, pattern recognition, machine learning [1], compression, classification and various disciplines as psychology, business, marketing, biology, libraries, insurance, city-planning and earthquake studies. [4, 5]

Many data clustering algorithms have been proposed in the literature. These algorithms can be classified into hierarchical clustering, partitional clustering algorithms, artificial neural networks for clustering, statistical clustering algorithms, density-based clustering algorithm, evolutionary approaches for clustering, search-based approaches and so on, [1, 2, 6, 7]. In these techniques, hierarchical and partitional clustering algorithms are the primitive approaches for data clustering. Hierarchical clustering algorithms can usually find pleasure clustering results. It is able to find different clustering results for different similarity or dissimilarity requirements.

The aim of this study is to determine the effects of the winds on the evaporation losses and to demonstrate whether according to the winds,(i.e. monthly mean wind speed and/or wind blow number) and evaporation loses, the months for each year can be clustered climatologically. For the aim, the data of (1) monthly mean wind speed

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

obtained from hourly mean wind speed, wind blow number measured for one day and the data of (2) monthly mean evaporation obtained from daily measured data, in the area of Eğirdir Lake have been used. Hierarchical clustering algorithm was used for clustering.

Hierarchical clustering algorithm

A clustering result established by a hierarchical clustering algorithm has a hierarchical structure. "The operation of a hierarchical clustering algorithm is illustrated using the two-dimensional data set in *Fig. 1*. This figure depicts seven patterns (or observation, datum or feature vector) labelled A, B, C, D, E, F, and G in three clusters. A hierarchical algorithm yields a *dendrogram* representing the nested grouping of patterns and similarity levels at which groupings change. A dendrogram corresponding to the seven points in *Fig. 1* (obtained from the single-link algorithm [Jain and Dubes, 1988]) is shown in *Fig. 2*. The dendrogram can be broken at different clustering of the data, [6]".With the hierarchical structure different clustering results can be obtained for different similarity requirements as shown in *Fig. 1*. For instance If the similarity requirement is set at level 1, the input dataset is partitioned into three clusters, i.e., {(A, B, C)}, {(D, E)} and {(F, G)}.

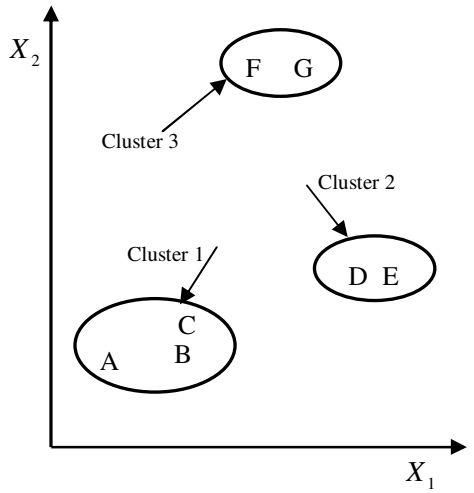

Figure 1. *Points fall in three clusters [6].*

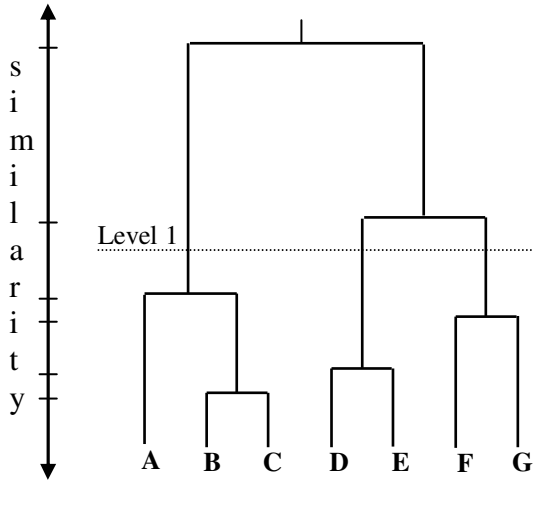

Figure 2*. The dendrogram obtained using single-link algorithm [6].*

Most of hierarchical clustering algorithms are variations of the single-link and complete-link algorithms. The both of them characterize the similarity between a pair of clusters in different way. In the first method the distance between two clusters is the minimum of the distances between all pairs of patterns obtained from two clusters. And in the other, the distance between two clusters is the maximal distance of all pair-wise of patterns in the two clusters. These algorithms are also explained superficially [7] and comprehensively in [6] studies.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

Similarity measures

A measure of the similarity between two patterns derived from the space, having same characteristic, is essential to most clustering procedures. Hence, similarity is fundamental for determining a cluster. The distance measure/measures must be chosen carefully due to the variety of feature types and scales. For calculating the dissimilarity between two patterns, the most popular method is to use a distance measure defined on the feature space.

The well-known distance measure, used for patterns of which features are all continuous, is the *Euclidean distance* which is a special case of: [8]

$$d_{ij} = d_m(x_i, x_j) = \left(\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^m \right)^{1/m} \qquad \text{(Eq.1)}$$

where; m varies in case of a distance measure to assign similarity. For m=1, 2 and 3, the eq.(1) gives the City block distance, the Euclidean distance and the Minkowski distance, respectively. The Euclidean distance is commonly used to evaluate the proximity of objects in two or three dimensional space.

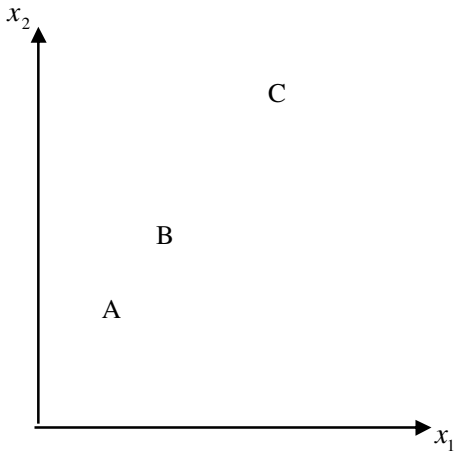

Figure 3. *A and B are more similar than A and C [4].*

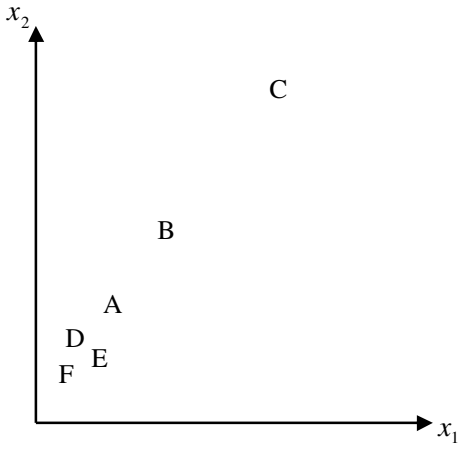

Figure 4. *After a change in context, B and C are more similar than B and A* [4].

There are some distance measures reported in Gowda and Krishna [6] that take into account the effect of surrounding or neighbouring points. The set of surrounding points is called *context*. A metric defined by using context is *the mutual neighbour distance* (MND), proposed in Gowda and Krishna [1977]. It is explained comprehensively in [6].

This measure is given by;

$$MND(x_i, x_j) = NN(x_i, x_j) + NN(x_j, x_i) \qquad \text{(Eq.2)}$$

where $NN(x_i, x_j)$ is the neighbour number of x_j with respect to x_i. *Fig. 3* and *4* give an example. In *Fig. 3*, the nearest neighbour of A is B, B's nearest neighbour is A. So, $NN(A,B) = NN(B,A) = 1$ and the MND between A and B is 2. However, $NN(B,C) = 1$

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

but $NN(C,B)=2$ and therefore $MND(B,C)=3$. *Fig. 4* was obtained from *Fig. 3* by adding three new points D, E, and F. Now $MND(B,C)=3$ (as before), but $MND(A,B)=5$. The MND between A and B has been increased by introducing additional points, even though A and B haven't been moved.

Available data

The data used in this study were obtained from the Regional State Hydraulic Works and State Meteorological Works' meteorological stations in the basin of the Lake area, situated about 917 meters above sea level on the south western part of the Mediterranean Region in Turkey, at about 30° 18′-31° 22′ eastern longitudes and 37° 48′-38° 26′ northern latitudes. It is one of the most important lake (the second largest freshwater) in Turkey and operated for multiple purposes [3, 9].

The wind climate of the region is normal but in summer, especially, north winds blow intensively. These effective winds blow through the south-north direction, which is the longer part of the Lake (i.e., it is about 50 km length). Also, in summer, evaporation losses gain more importance with respect to the reservoir of the Lake and water demand.

Based on the records of the wind gauge station and Class-A evaporation pans situated near the Lake, the used average monthly data sets were obtained from the records between 1930-1999. For simplicity, the monthly evaporation loss was directly obtained from the daily observation data, by multiplying the measured rate by the pan coefficient.

Application

In this study, as a nonparametric approach hierarchical clustering algorithm is applied in clustering analysis of the monthly evaporation losses with the monthly mean wind speeds and wind blow numbers of the Lake Egirdir. In the clustering, the similarity is determined by the mutual neighbour distance (MND) algorithm.

The winds have two important characteristics, increase the lateral transport as well as turbulent diffusion in the vertical direction and therefore have an important effect on the evaporation rate. These are speed and direction. The first characteristic has been represented by numerical values such as monthly mean speeds. Because of the lack of monthly mean values of directions, practically, the effective wind direction can be chosen for the general wind direction to determine the effect of the winds on the evaporations. In the application; (1) the wind speed and evaporation data sets (2) wind blow number and evaporation data sets were clustered in different stages. And for the different similarity levels, the relationships between winds and evaporation loss were illustrated.

The hierarchical clustering algorithm is applied to three observed dataset which their mean values are given in *Table 1*. One of these is the wind speed, the second is the wind blow number and the last one is the evaporation losses. In the first clustering analysis the wind speed and in the second analysis the wind blow number was accepted as a first dimension and the evaporation loss was chosen as a second dimension in both of analysis.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

Table 1. *The mean values of the data sets*

Months	Oct (1)	Nov (2)	Dec (3)	Jan (4)	Feb (5)	Mar (6)	Apr (7)	May (8)	June (9)	July (10)	Aug (11)	Sept (12)
Wind speed (m/sn)	2,4	3,6	3	3,2	3,5	3,3	3,3	2,9	3,2	3	2,9	2,7
Wind blow number	407	375	388	438	364	422	319	424	577	799	732	599
Evaporation (mm)	112.1	48,3	0	0	0	35	124	182,8	235,9	305,5	287	200

Table 2. *Proximity matrixes of (a) wind speed-evaporation, (b) wind blow number-evaporation(a)*

	Euclidean distance											
Months	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0,00	63,86	112,12	112,12	112,13	77,13	11,93	70,67	123,79	193,42	174,9	87,94
2	---	0,00	48,27	48,27	48,27	13,27	75,75	134,52	187,64	257,27	238,75	151,79
3	---	---	0,00	0,20	0,50	35,00	124,02	182,79	235,91	305,54	287,02	200,06
4	---	---	---	0,00	0,30	35,00	124,02	182,79	235,91	305,54	287,02	200,06
5	---	---	---	---	0,00	35,00	124,02	182,79	235,91	305,54	287,02	200,06
6	---	---	---	---	---	0,00	89,02	147,79	200,91	270,54	252,02	165,06
7	---	---	---	---	---	---	0,00	58,77	111,89	181,52	163,00	76,04
8	---	---	---	---	---	---	---	0,00	53,12	122,75	104,23	12,27
9	---	---	---	---	---	---	---	---	0,00	69,63	51,11	35,85
10	---	---	---	---	---	---	---	---	---	0,00	18,52	105,48
11	---	---	---	---	---	---	---	---	---	---	0,00	86,96
12	---	---	---	---	---	---	---	---	---	---	---	0,00

(b)

	Euclidean distance											
Months	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0,00	71,42	113,72	116,33	120,08	78,57	88,80	72,69	210,30	437,12	369,07	221,18
2	---	0,00	50,00	79,37	49,51	48,84	94,20	143,17	275,70	495,95	429,48	270,59
3	---	---	0,00	50,00	24,00	48,80	141,92	186,30	302,28	512,13	448,01	290,77
4	---	---	---	0,00	74,00	38,48	171,88	183,33	273,82	472,94	410,87	256,80
5	---	---	---	---	0,00	67,74	131,93	192,39	317,84	531,58	466,70	308,62
6	---	---	---	---	---	0,00	136,14	147,80	253,75	464,03	399,52	242,02
7	---	---	---	---	---	---	0,00	120,33	281,22	513,18	444,00	290,14
8	---	---	---	---	---	---	---	0,00	161,96	394,58	325,16	175,85
9	---	---	---	---	---	---	---	---	0,00	232,66	163,21	42,06
10	---	---	---	---	---	---	---	---	---	0,00	69,51	226,11
11	---	---	---	---	---	---	---	---	---	---	0,00	158,91
12	---	---	---	---	---	---	---	---	---	---	---	0,00

In both of applications, the values of these dimensions are plotted on x and y axis. The distance between each point in both dataset is calculated by using the Euclidean distance, a special case of equation 1. The distance measures, showed in the proximity matrixes of the wind speeds-evaporation losses and wind blow number-evaporation losses with 12x12 dimensions (*Table 2.*), are ordered from the smallest to highest. Hence, the neighbour degrees of each dual point $[ND[N_i, N_j]]$, represent dual months, are determined. It can be clearly perceived that the nearest neighbour is itself for each

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

point. So, $ND(N_i, N_i) = 0$ and the neighbour degree of the furthest point from the current point will be 11 subject to the number of proximity objects or the surrounding points. Relatively, for each 12 points, the total neighbour number of the dual points $N(ND)$ is equal to 144.

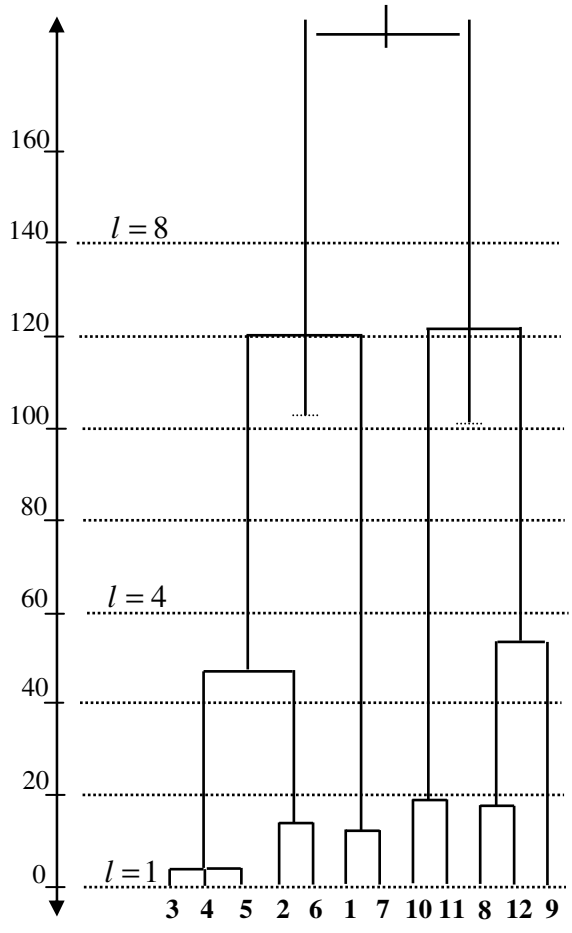

Figure 5. *The dendrogram depends on hierarchical single-linkage for the first application.*

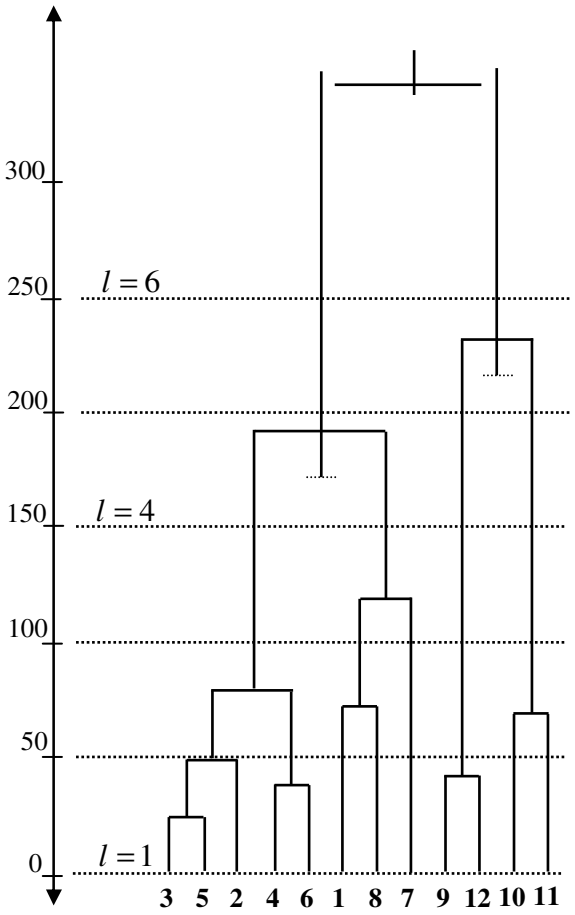

Figure 6. *The dendrogram depends on hierarchical single-linkage for the second application.*

To make easy to generate the dendrogram, the neighbour degree of each point with itself is taken into account as 1 and the *MND* between x_i and x_j is calculated by summation of $ND(N_i, N_j)$ and $ND(N_j, N_i)$. Also, the *MND* is the sum of the number of the combinations of the different proximity objects in which there are sub-clusters. In both of applications or surveys, the *MND s* are 78, (i.e. $C\binom{12}{1} + C\binom{12}{2} = 78$). The dendrogram, derived from the proximity matrix that emphasizes the distance of a point from the other points, for each group, depends on hierarchical single linkage, were generated independently and given in *Figs. 5* and *6*. In these dendrograms, the dataset can be grouped for different similarity levels.

The similarity level is represented by $S_l ; l = 1,2,3,...,m$. As it can be seen from both *Figs. 5* and *6*, the dendrogram can be broken at any similarity level to yield different

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

clustering results. For wind speed-evaporation the similarity level (l) varies from 1 to 9 and for wind blow number-evaporation it varies from 1 to 7. S_1 means that there is no cluster. In other words, each cluster has only one pattern. For instance, for the wind speed-evaporation dendrogram there are 4 and 2 clusters at S_4 and S_8, respectively, whereas these clusters can be seen at S_4 and S_6 for the wind blow number-evaporation dendrogram. For S_7 and S_9 in both dataset there is only one cluster. This means that all the patterns belong to only one cluster. The clusters of various similarity levels for the first ($l = 4 \, and \, 8$) and second ($l = 4 \, and \, 6$) analysis, are shown in *Fig. 7* and *Fig. 8*.

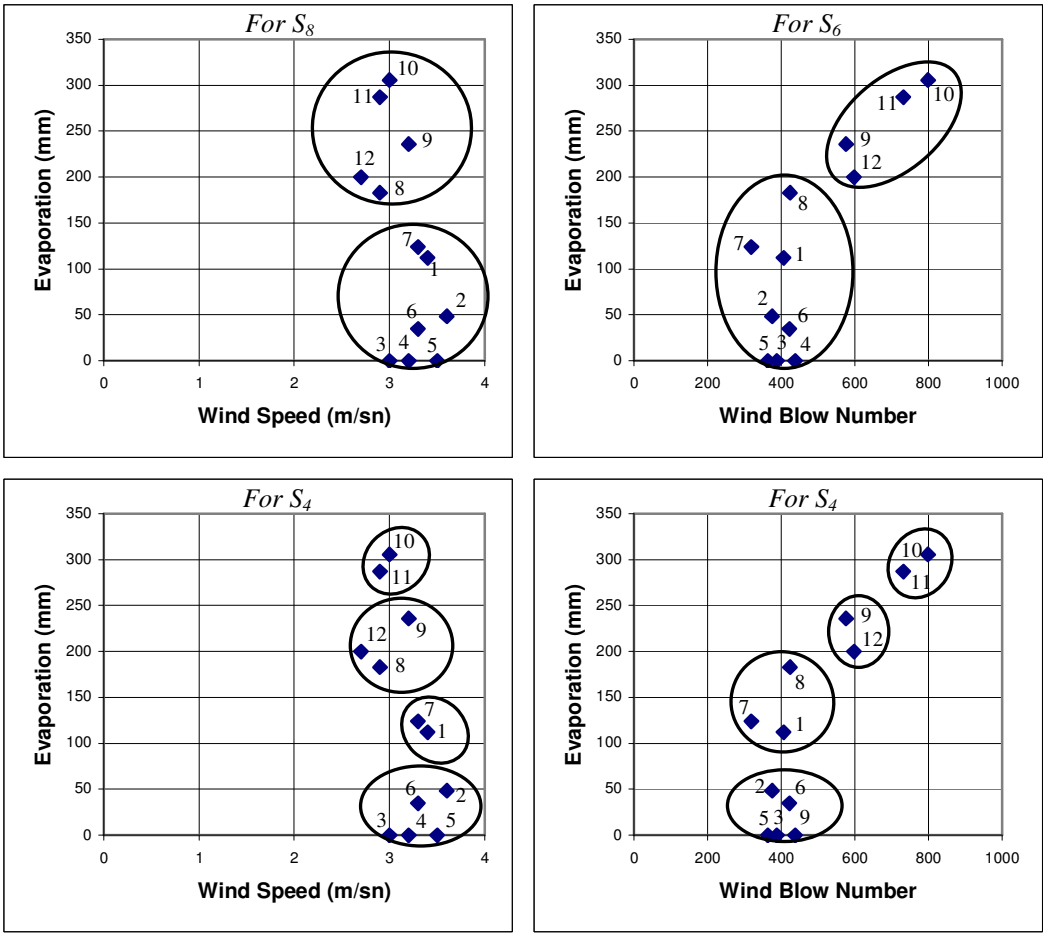

Figure 7. *The cluster for* $l = 4$ *and* $l = 8$.

Figure 8. *The cluster for* $l = 4$ *and* $l = 6$.

Result and conclusion

It is known that, in general, any change in the wind speed and blow number causes the change in evaporation rate when the other meteorological factors, effecting on the evaporation rate, are assumed as constant. In this study, effects of the winds on the evaporation losses were investigated. The following interpretations were obtained.

(1) It can be considered that the clustering at S_1 (i.e. all the patterns belong to only one cluster.) for the first data set, shows that there cannot be seen any statistical

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

relationship between the change in the wind speed and the change in the evaporation amount ($R^2 = 0{,}02$), however, for the second data set, the same similarity level indicates that the effects of winds on the evaporation rate ($R^2 = 0{,}69$) has a quite important effect, statistically.

(2) Although the first derived explanation, it is thought that the clustering for a second data set from the same data source, for instance; at S_8 ($l = 8$) similarity level for wind speed-evaporation rate, the cluster which contains July, August, Jun, May and September and at S_6 ($l = 6$) similarity level for the wind blow number-evaporation rate, the cluster which contains July, August, Jun, and September; state the above relations strongly ($R^2 = 0{,}29$ for wind speed change and evaporation rate; $R^2 = 0{,}85$ for the wind blow number-evaporation rate).

(3) It can be possible that making some critical analyses by clustering method for resolving the problems such as there can't be seen clearly a correlation between two or more patterns selected for examination related to the problem analysis. For instance, in this study when looking into the wind blow number-evaporation rate data set it cannot be asserted any statistical statement exactly about the relationship between these two patterns however, by clustering at $l = 6$ (the cluster which contains June, July, August, September) a strong relationship can be easily stated ($R^2 = 0{,}96$).

(4) It is considered that the clustering should be used in determining of the different operation level for each similarity level or in making efficient operating decisions and making accurate prediction.

(5) In general, in order to introduce the determination of a stress relation between anyhow two objects or patterns, having a scientific or statistic meaning, the clustering method presents that what data types and groups represents the objects in the best way.

REFERENCES

[1] BERKHIN, P. (2005): Survey of Clustering Data Mining Techniques – Accrue Software, Inc. http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf

[2] BUHMANN, J.M. (2002): Data Clustering and Learning – Handbook of Brain Theory and Neural Networks, M. Arbib (ed.), 2nd edition, MIT Press

[3] DSİ, Eğirdir Gölü Hidrolojisi, Revize Raporu, T.C. ETKB-DSİ Gen. Müd., 18.Bölge Müdürlüğü, Isparta, 2002.

[4] FUNG, G. (2001): A Comprehensive Overview of Basic Clustering Algorithms. – http://www.cs.wisc.edu/~gfung/clustering.pdf

[5] JAIN, A.K., MURTY, M.N and FLYNN P.J. (1999): Data Clustering: A Review – ACM Computing Surveys, Vol. 31, No. 3,

[6] JAIN, A.K. and DUBES, R.C. (1988): Algorithms for Clustering Data. – Prentice-Hall advanced reference series. Prentice-Hall, Inc.,Upper Saddle River,NJ.

[7] LİN, Cheng-Ru and CHEN, Ming-Syan (2005): Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging – IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 2.

[8] LIONEL, E., et al. (2004): Efficient Clustering-Based Genetic Algorithms in Chemical Kinetic Modelling – K. Deb et al. (Eds.): GECCO, LNCS 3103, pp. 932–944.

[9] Utku, M. (1990): Isparta İklim Etüdü, DMİ Arş. Bil. İşlem D. Baş. Arş. Şb. Müd., ISP.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623