Journal of Engineering and Applied Sciences 2 (3): 524-530, 2007

A Comparison of Low-Flow Clustering Methods: Streamflow Grouping

[1]Ercan Kahya and [2]M. Cüneyd Demirel
[1]Department of Civil Engineering, Istanbul Technical University,
Maslak, 34469 Istanbul, Turkey
[2]Institute of Science and Technology, Istanbul Technical University,
34469 Maslak, Istanbul, Turkey

Abstract: In this study, three clustering algorithms which use agglomerative clustering procedure to identify groups of similar catchments are investigated to determine their effectiveness in low flow clustering scheme. These hierarchical clustering algorithms are single linkage, complete linkage and Ward's algorithms. The effectiveness of the cluster analysis algorithms is investigated by using monthly minimum streamflow data recorded from watersheds in Turkey with the period of 1964-1994. Furthermore one of cluster validity indices (namely, cophenet correlation coefficient index) is used to strengthen our results. The hierarchical cluster analysis is found to be useful in minimizing efforts needed to identify homogeneous clusters. Ward's algorithm with Euclidean metric is the one to decrease the variance in each group and appears to be favourable in the applications of low-flow. The results and comparison with our previous studies are also presented in the thematic maps.

Key words: Cluster analysis, low-flow, homogeneous region, Ward's method, Turkey, thematic maps

INTRODUCTION

Cluster analysis algorithms are mostly used in hydrology for regional analysis of floods, low flows, rainfall and other river basin variables. If the clustering scheme is successful, strong relationships between streamflow properties (e.g., mean, standard deviation and correlation of monthly flows) can be achieved. Then these relationships can be used to develop the streamflow information at ungauged watersheds with similar patterns. Although, in the last two decades, clustering methodology in hydrology was improved by many efforts, no single procedure has been demonstrated to yield universally acceptable results yet. Several types of hierarchical clustering methods and similarity metrics are available and to make different combination of similarity metric-clustering algorithm increases the solution possibility.

In hydrology, continuous demands for irrigation water and hydropower, flood control, municipal and industrial water supplies and soil conservation issues led to major activity in extreme flow analysis (Riggs, 1985). On the other hand, the water deficit situation causing unusual droughts or yearly recurring low-flow increased the importance of drought analysis and low-flow studies. Stahl (2001) studied drought across Europe by correlating the monthly averages of the Regional Streamflow Deficiency Index (RDI) series of the 19 European clusters to the NAO indexes and found weak correlations. Most rivers in Europe domain show a strong seasonal regime; therefore, seasonal variability was important to assess the impact of climate changes on this complex hydrological system (Stalh, 2001). Dettinger and Diaz (2000) worked with the monthly streamflow series in global scale and indicated that the timing and amplitude of seasonality in streamflow depend on the local month of maximum precipitation and the extent to which precipitation is trapped in snow and ice at most gauges. In the same context, we defined the main objective of this study as exploring the geographical zones having the similar low-flow patterns.

MATERIALS AND METHODS

Data: The study domain is the entire Turkey, located in a semiarid zone where precipitation is mainly characterized by high spatial and temporal variability. Readers are referred to Ünal *et al.* (2003) and Karaca *et al.* (2000) for a recent review of the general climate features of Turkey. Each streamflow station contains a 31-year period spanning from 1964 to 1994. Minimum monthly streamflow records at 80 stations across Turkey compiled by General Directorate of Electrical Power Resources Survey and Development Administration (abbreviated as EIE) are used in this study. The data used in this study was previously shown to satisfy the homogeneity condition at a desirable confidence (Kahya and Karabork, 2001).

Corresponding Author: Ercan Kahya, Istanbul Technical University, Department of Civil Engineering, Maslak, 34469 Istanbul, Turkey

Preliminary analysis: In this study, all the data pre-processing and cluster analysis scheme were carried out for two cases: Case I is namely used for monthly minimum streamflow data resulted from three algorithms (Single Linkage, Complete Linkage, Ward) and Case II is used to indicate the same algorithms with standardized monthly minimum streamflow data. Standardization was done by z-score values. Multidimensional Scaling (MDS) was applied to the data for screening their weights in two dimensional coordinates (not shown here). There is a high accumulation in Case I when the raw data is used; hence, the use of standardized, equally weighted data is superior for the application. Since the general view of the screening is almost similar in three algorithm results, only MDS of Ward method was given to demonstrate importance of standardization. This issue was also emphasized by Demirel (2004) and standardization by range was referred as the best performance among others. Without pre-information by applying MDS, one can also visualize the irrelevant separation and accumulation in one cluster from the dendrogram of raw data (referred as Case I). Figure (1-8) are given here.

Cluster analysis: The main purpose of cluster analysis is to organize observed data into meaningful structures with any priori knowledge. In a successful scheme, the clusters must have the properties of internal cohesion and external isolation (Everitt, 1993). The similarity metric between two objects i and j, the Euclidean distance function, is frequently used (Chiang, 1996; Gong and Rchman, 1995):

$$d_{ij} = \sum_{i=1}^{n} (x_{ik} - x_{kj})^2 \qquad (1)$$

where p is the number of variables. The majority of investigators (85%) applied this metric in their study (Gong and Richman, 1995). The agglomerative hierarchical clustering methods are considered to be the most popular cluster analysis technique (Gong and Richman, 1995). A combination of Euclidean metric and three linkage methods was here used to group streamflow stations in a multi-dimensional space in order to discover to natural structure of low-flow patterns in Turkey. Ward's algorithm (1963) and Euclidean distance were selected in this analysis as this linkage method minimizes the variance within a cluster. Hence two clusters are merged if this merger results in a minimum increase in the Error Sum of Squares (ESS) (Everitt, 1993).

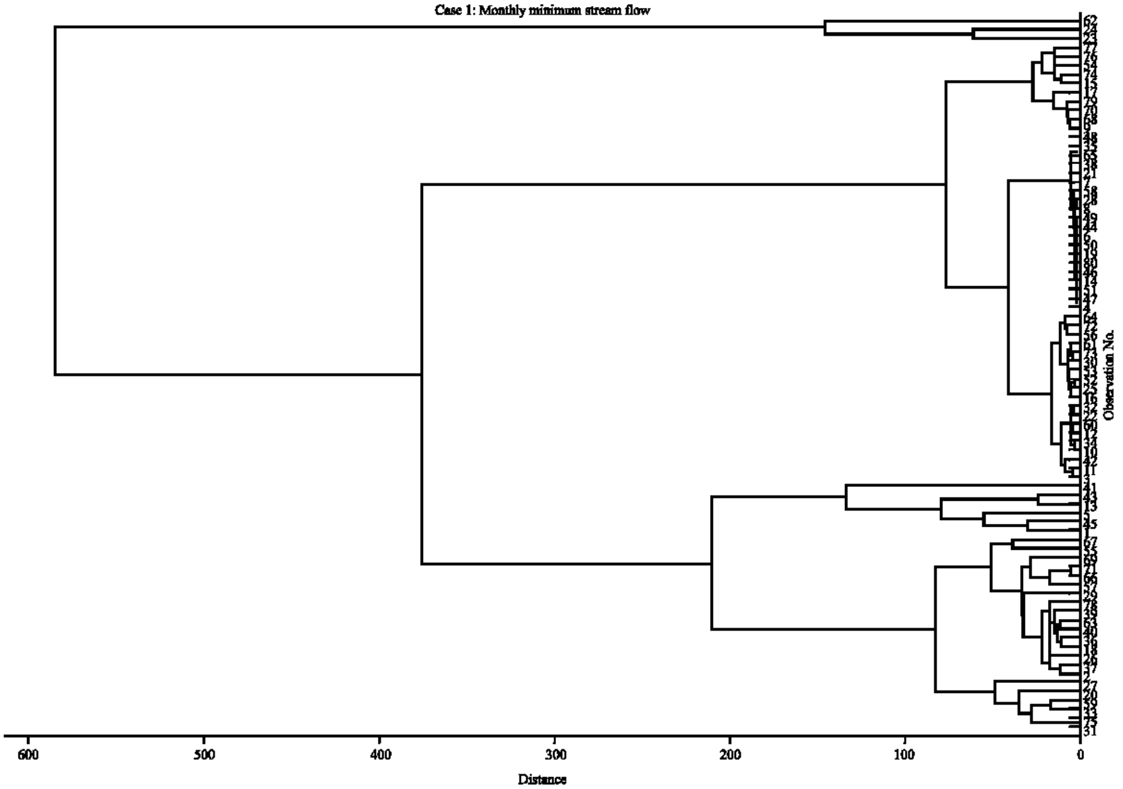

Fig. 1: Case I dendrogram of the Euclidean distance-Ward method combination

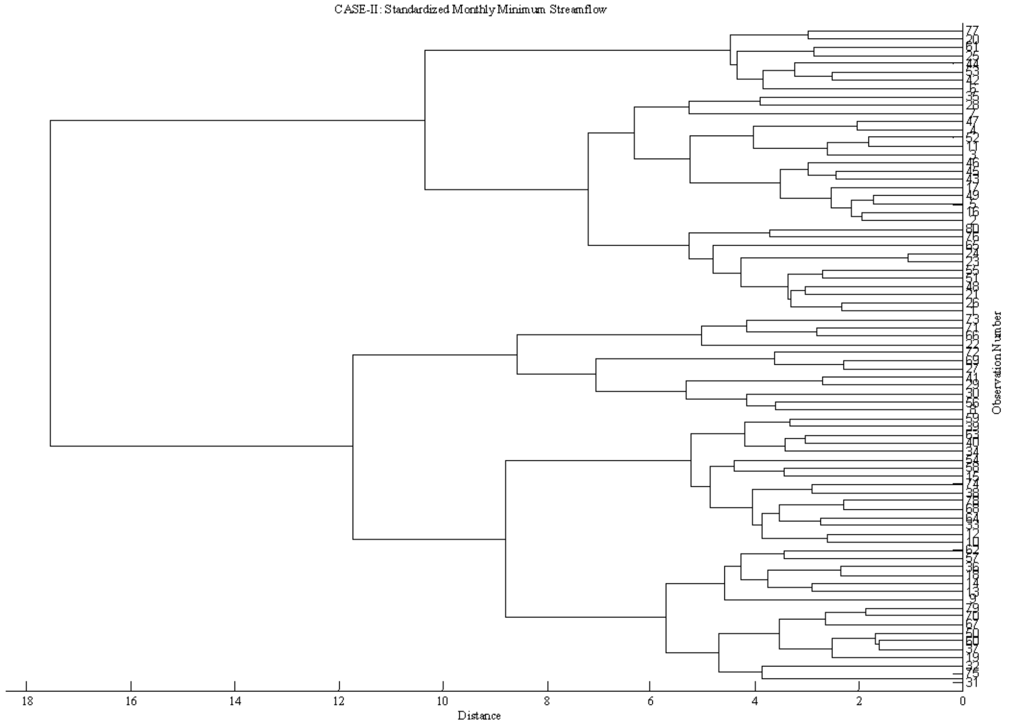

Fig. 2: Case II dendrogram of the Euclidean distance-Ward method combination

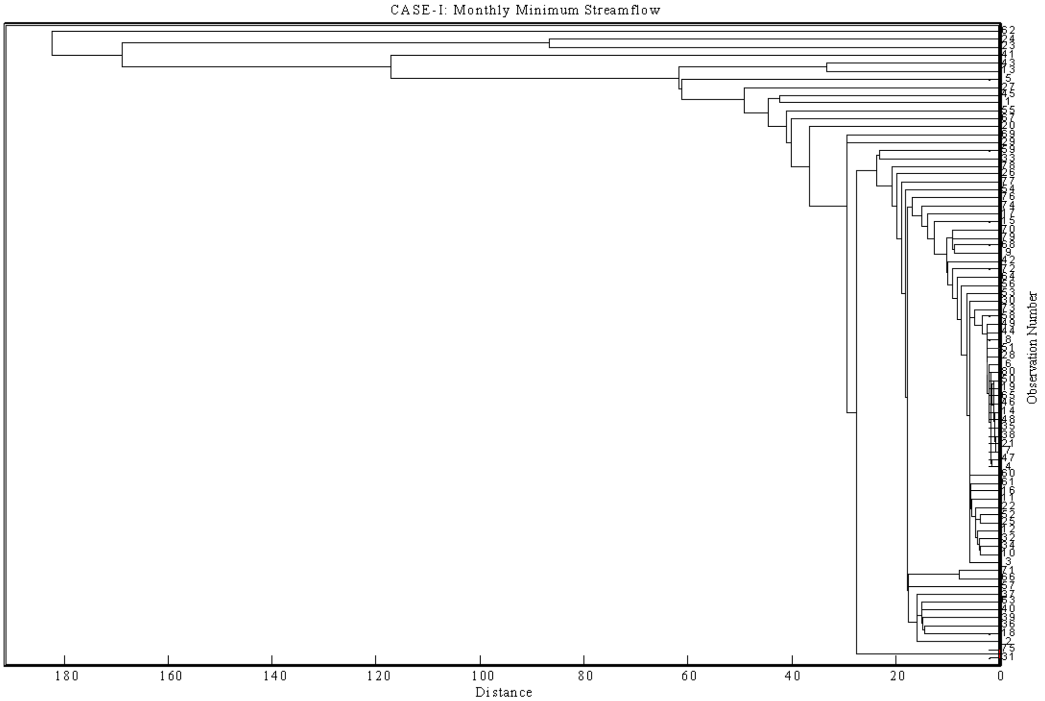

Fig. 3: Case I dendrogram of the Euclidean distance-Single Linkage (SL) method combination

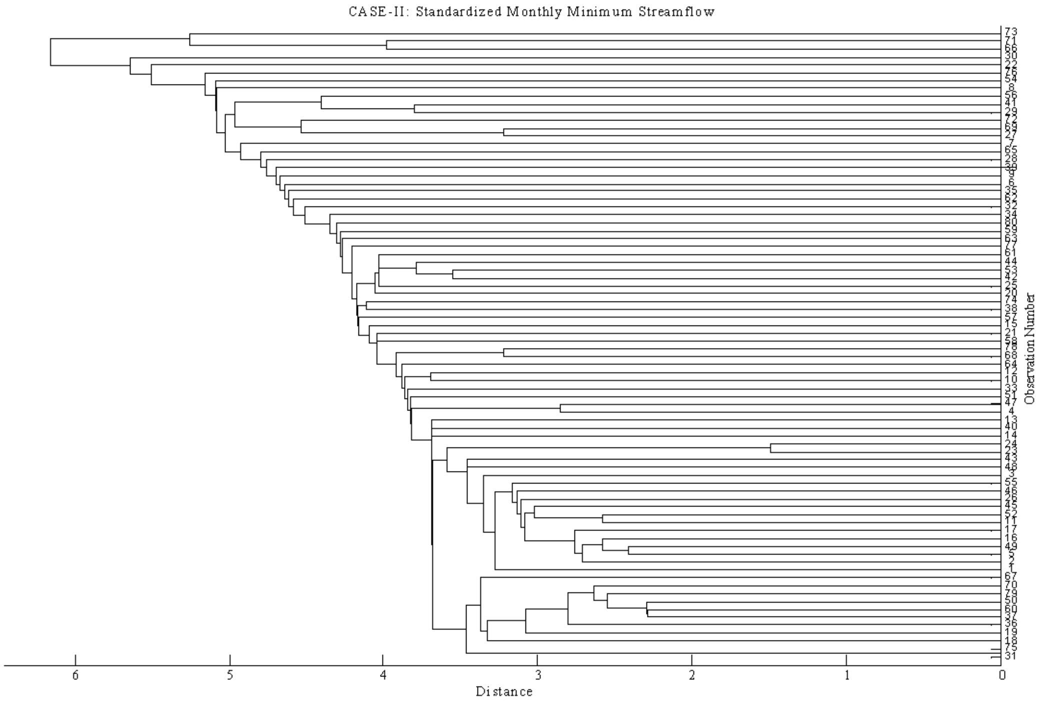

Fig. 4: Case II dendrogram of the Euclidean distance-SL method combination

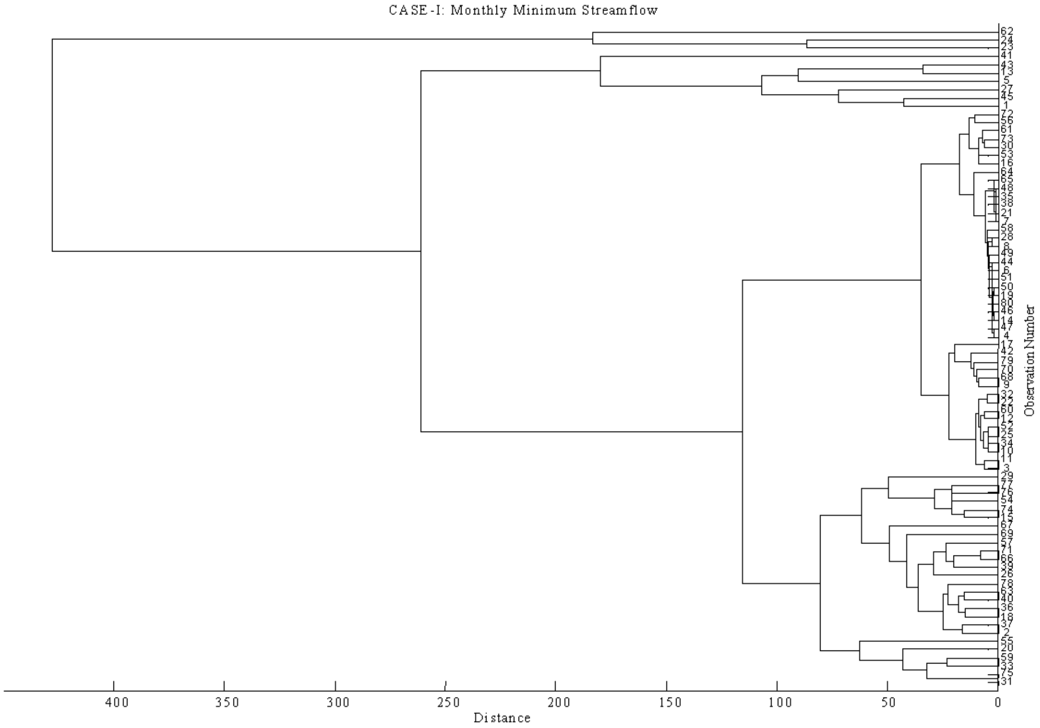

Fig. 5: Case I dendrogram of the Euclidean distance-CL method combination

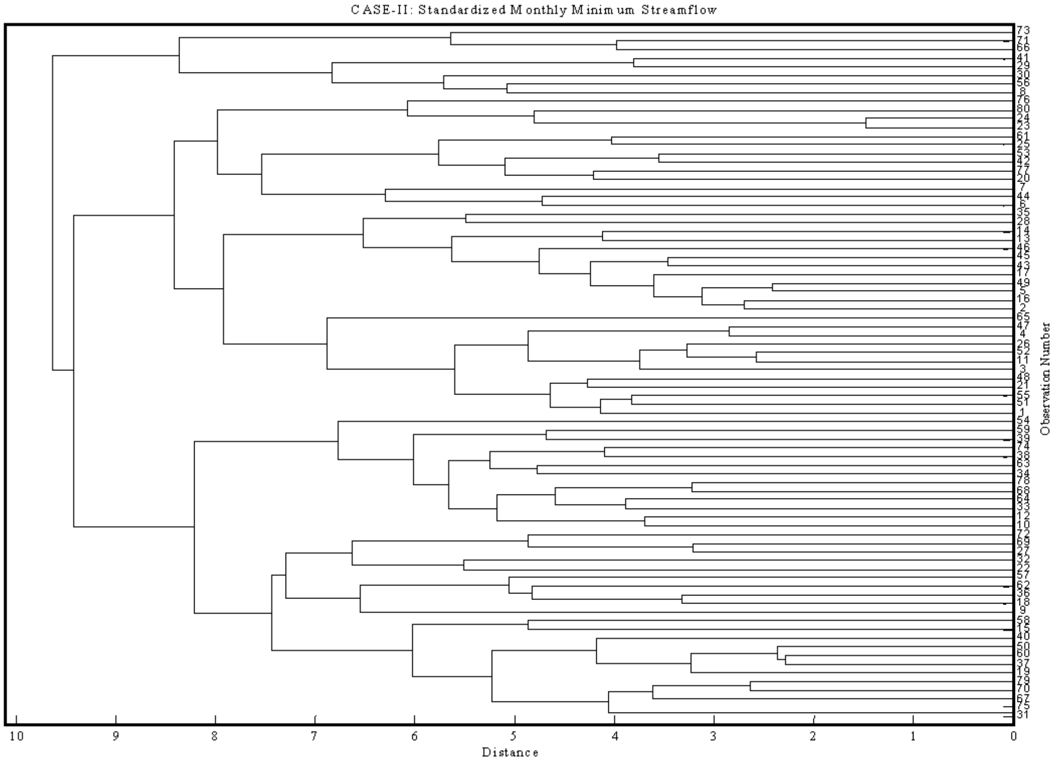

Fig. 6: Case II dendrogram of the Euclidean distance-CL method combination

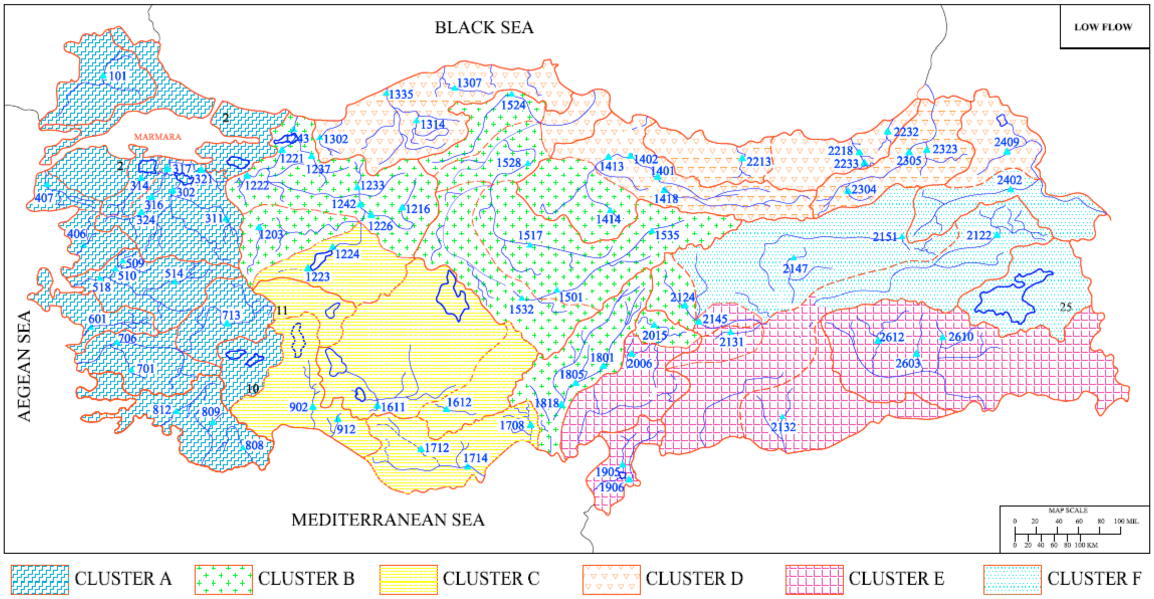

Fig. 7: Low-flow clusters obtained by Ward's algorithm with Euclidean metric

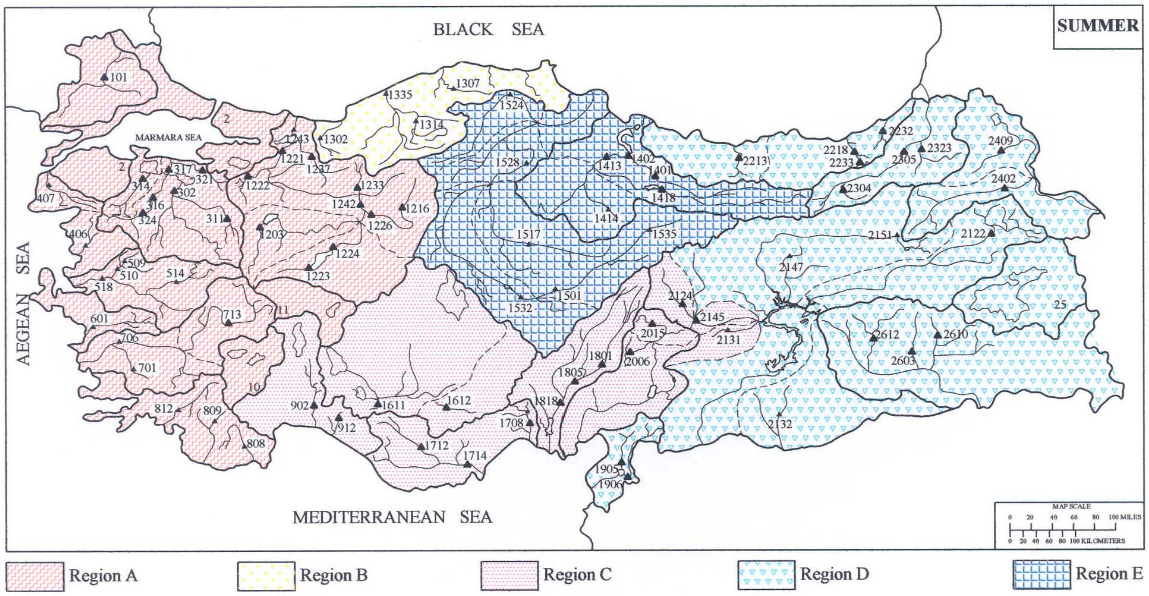

Fig. 8: Summer season clusters, adapted from Demirel (2004)

$$ESS = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \qquad (2)$$

The dendrogram helps user examine visually the resultant structure; however, the optimum number of clusters to retain is an issue in the cluster analysis so that the use of various tests having advisory role is recommended to decide where the clustering procedure should be stopped (Gong and Richman, 1995).

RESULTS AND DISCUSSION

Three linkage methods with Euclidean metric were used to group 80 streamflow stations in Turkey. We selected 6 as the number of groups based on examining the dendrograms and our previous study (Demiral, 2004). In fact, different number of clusters could be defined by moving a virtual line perpendicular to axis in the dendrogram, left or right to achieve the structure in desired detail level (i.e., 3, 5 and 8) (01). Case I scenarios was not taken into account in mapping; nevertheless, they emerged higher cophenet coefficient values regarding the distortion of the clustering, indicating how readily the data fits into the structure suggested by the grouping (MATLAB HELP). It is worth noting that the dendrogram structures for Case I are not readable to create an equally weighted low-flow maps (i.e., 01, 03 and 05). This is a clear indication that could also be reached after applying MDS process before clustering scheme. The high values for Single Linkage (SL) and Complete Linkage (CL) can be related to chaining problem of these two methods in the literature (Everitt, 1993). Both 03 and 04 show the chaining problem by the use of single linkage algorithm in the two cases.

Table 1: Performance results of the Euclidean metric and three linkage methods

Method Combinations	Cophenet coefficient Case I	Case II
Euclidean and single linkage	0.95411	0.65492
Euclidean and complete linkage	0.92468	0.64783
Euclidean and ward	0.75243	0.57928

Case II scenarios resulted in better solutions, except SL algorithm, according to their dendrogram results with equally weighted groups (0 and 06). Ward method is subjected to create lowest variance value for each group; thus, Case II with Ward algorithm was chosen for the thematic map regarding the low flow map Turkey (Everitt, 1993). The highest value of cophenet coefficient index was reached with SL algorithm in Case I (Table 1). The values closer to unity indicate better solutions so that the value of 0.95411 (Case I-SL) should be the best scheme; however, in this solution, dendrogram shows the chaining effect clearly (03). Similarly, the value of 0.65492 is the highest value for the second column (Case II) but not the correct clustering plan (04).

Highly accumulation in one cluster is not desired in clustering applications; hence, the scale difference or possible noises from the data should be excluded. The grid technique, which makes the study domain having more or less uniform station distribution, can be applied as a data pre-processing tool to avoid the detrimental effects of inequality in network density. In this study, the resultant summer seasonal cluster analysis map prepared by Demirel (2004) also included to show the relations with cluster A-region A indicating a large western group. Region C in 0 and cluster C in 0 have many stations in common based on a relation with low-flow and summer season clustering results. The complete linkage in Case I also failed to show equal distribution; that is to say, accumulation problem like other Case I clustering scheme due to raw data effect (07). However the complete linkage in Case II revealed relatively better distribution of stations in each cluster (06).

The coastal river basins, mainly the eastern black sea and the south eastern river basin stations show a strong cohesion that makes them always in the same cluster (07).Aegean region left as western dominating block in many time domain analysis of Demirel (2004). The cluster clearly defined as cluster C is located in the mid-south parts of the country where low-flow conditions are mainly dominant over Konya basin whose mean annual precipitation total is between 416.8 and 474.3 mm (SIS, 1995). The previously defined clustering level for summer season was only five (Demirel, 2004) (08). In this study, six clusters were found to be sufficient to represent the low-flow clustering pattern over the study domain(07).

CONCLUSION

The physical processes of low-flow (i.e., river basin characteristics, climatology, etc.) must be taken into consideration to have a better clustering scheme in hydrological perspective. The use of monthly patterns was more favourable for defining low-flow homogeneous regions. Standardization was inevitable to get equally weighted clusters. The use of z-scores performs well in the robust clustering scheme. The Ward's method with Euclidean was more effective in producing homogenous clusters comparing to single linkage and complete linkage methods. Cophenet coefficient index couldn't be effectively used as the validity index to choose better method due to chaining effects of single linkage method in both cases. However the replication analysis by dividing the data set into two parts is also necessary to get robust cluster structure and to avoid stability concerns. The grid lines on the map to choose the equal number of stations representing the river basins is suggested to apply for station reduction.

ACKNOWLEDGEMENT

This research is supported by the Istanbul Technical University (ITU) Research Activities Secretariat.

REFERENCES

Riggs, H.C., 1985. Streamflow characteristics. Elsevier, Amsterdam.

Stahl, K., 2001. Hydrological drought: A study across Europe. Ph.D Thesis, Albert-Ludwigs-Universität, Freiburg.

Dettinger, M.D. and H.F. Diaz, 2000. Global characteristics of streamflow and variability. J. Hydrometeorol., 1: 289-310.

Ünal, Y., T. Kindap and M. Karaca, 2003. Redefining the climate zones of Turkey using cluster analysis. Int. J. Climatol., 23: 1045-1055.

Karaca, M., A. Deniz and M. Tayanc, 2000. Cyclone track variability over Turkey in association with regional climate. Int. J. Climatol., 20: 122-136.

Kahya, E. and M.Ç. Karabörk, 2001. The analysis of El Nino and La Nina signals in streamflows of Turkey. Int. J. Climatol., 21: 1231-1250.

Demirel, M.C., 2004. Cluster Analysis of Streamflow Data over Turkey. M.Sc Thesis, Istanbul Technical University, Istanbul.

Chiang, S.M., 1996. Hydrologic regionalization for the estimation of streamflow at ungauged sites based on time series analysis and multivariate statistical analysis. Ph.D Thesis, Syracuse University, New York.

Gong, X. and M.B. Richman, 1995. On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. J. Climate, 8: 897-931.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc., 58: 236-244.

Everitt, B., 1993. Cluster analysis. Halsted Press, Division of Wiley, New York.

State Institute of Statistics, 1995. Environmental statistics and river basin statistics. Publication Ankara, 2152: 120.