Performing k-means analysis to drought principal components of Turkish rivers

M. Cüneyd Demirel
Istanbul Technical University, Institute of Science and Technology, 34469 Maslak Istanbul, Turkey; also at Rosenstiel School of Marine and Atmospheric Sciences, Division of Meteorology and Physical Oceanography (RSMAS/MPO), University of Miami, Miami, Florida, USA

Arthur J. Mariano
Rosenstiel School of Marine and Atmospheric Sciences, Division of Meteorology and Physical Oceanography (RSMAS/MPO), University of Miami, Miami, Florida, USA

Ercan Kahya[1]
Istanbul Technical University, Civil Engineering Department, Hydraulic Division, 34469 Maslak Istanbul, Turkey

Abstract. In this study, the principal component analysis (here after PCA) was applied to 31-year (1964-1994) monthly minimum streamflow data from 23 catchments in Turkey. Ephemeral flows in winter are associated with non melted snow or even ice and summer low flows are related to the semi-arid climate of Turkey with topography, leading high temperature in lowland catchments. The PC matrix (80x2), explaining the highest variance of the main data, was chosen as an input to k-means routine to define drought regions. The first 4 PCs explains more than 80% of the total variance, the first PC presents 52.44% by itself. The resulting maps and silhouette plots for two level (6 and 10 clusters) scheme reveal that the clustering scheme is not successful when the principal components are used for defining the drought zones of Turkey.

1. Introduction

The use of multivariate techniques in hydro-climatological sciences has shed light on many climatological problems (i.e., defining the leading pattern). Stahl and Demuth (1999) applied a cluster analysis on the derived historical series of daily Regional Streamflow Deficiency Index (RDI) of the European domain to group into 19 regions, which are homogeneous in terms of simultaneous streamflow deficiency between 1962 and 1990. Stahl (2001) studied drought across Europe by correlating the monthly averages of the RDI series of these19 large clusters to the NAO indexes and found weak correlations. Nathan and McMahon (1990) applied different approaches to hydrological regionalization, which were based on a combination of cluster analysis, multiple regressions, PCA and the multidimensional scaling of data. The geographical continuity of homogeneous catchment groups is usually not observed in the resultant clusters (Smakhtin 2001; Demirel 2004). The identification of homo-

[1] Hydraulic Division
Civil Engineering Department
Istanbul Technical University
34469 Maslak Istanbul, Turkey
Tel: (212) 285-3002
e-mail: kahyae@itu.edu.tr

geneous regions is normally required for large domains such as continental based studies or areas with varying physiographic conditions. It may be ignored for smaller regions; however, highly sophisticated statistical techniques may not necessarily result in a more meaningful and practically applicable set of pattern groups than those administrative boundaries (Smakhtin, 2001).

This paper attempts to show the use of PCA together with k-means analysis over a country scale. The analysis will be carried out for two different numbers (levels) of clusters (6 and 10) to follow the change in cluster density. The scattering plot of cluster memberships and two silhouette diagrams of each level will be presented.

2. Description of study area

The study area covers the entire country and extends from 26-45° of longitude east and 36-42° of latitude north (Figure 1). The spatial distributions of the 80 continuous-record streamflow gauging stations are not uniform; however the monthly streamflow records compiled by EIE (General Directorate of Electrical Power Resources Survey and Development Administration) were shown to satisfy the homogeneity condition at a desirable confidence by Kahya and Karabörk (2001).

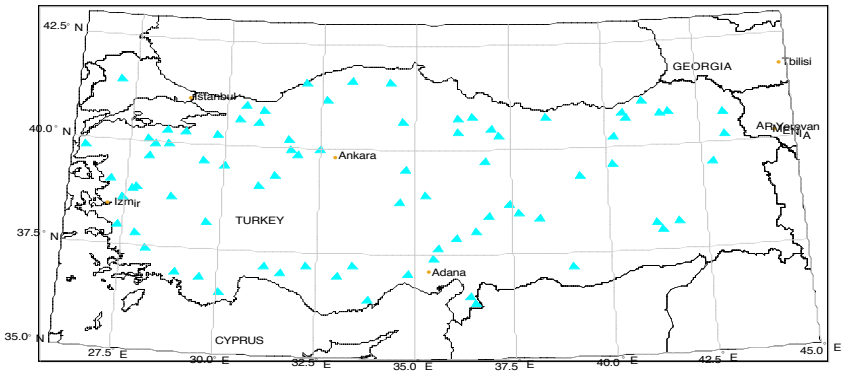

Figure 1. The streamflow stations used in the analysis.

3. Data

The standardized 31-year (1964-1994) monthly minimum streamflow data was used to achieve physically meaningful classifications, but it was also a useful preliminary analysis to avoid high scale value perturbation on others. This data from 80 stations covering 23 catchments in Turkey were used to develop PCA and k-means hybrid drought clustering scheme. The regulated streams were not included to the dataset.

4. Methods

The standardization by the range method applied to original streamflow values before using multivariate scheme (Demirel, 2004). The PCA and k-means analysis combination is used to identify zones with similar drought patterns so that the hydrological effects can be compared in these subregions, hydrologic predictions, transferring information from one area to another with

analogous characteristics can be possible if a robust scheme of regionalization is established (Andrade, 1997).

The main steps in k-means algorithm are as follows (Url-1):

1. Select an initial partition and define the centers.
2. Assign each entity (station) to the cluster that has the closest centre.
3. When all points have been assigned to one cluster, reorder the positions of the centers.
4. Repeat Steps 2 and 3 until cluster membership does not change.

Finally, this algorithm minimizes the *objective function*; in this case, a squared error function can be expressed as (MacQueen, 1967)

Euclidean distance: $d_{ij} = \left[\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right]^{1/2}$ (1)

The squared Euclidean is used as the measure of distance between an entity in the cluster and its respective cluster centre.

For the details concerning cluster analysis, readers are referred to Bacher (2002) who provided comprehensive expressions on this subject. The reference of Krasovskaia and Gottschalk (1995) is suggested for the detailed explanations of PCA method associated with clustering scheme.

5. PCA and K-means Clustering Results

We applied the k-means method to the first two PCs of the streamflow data which explain more than %70 of the total variation (Figure 2). The plot of these two PCs scores showed accumulation at an arbitrary point (Figure 3). The silhouette plots also provided to identify the heterogeneous spread (Figure 4).

The cluster 5, colored as purplish in Figure 5 (for the 6 cluster solution), is the largest cluster covering most of the country and does not reveal any interpretable meaning from the standpoint of hydrology.

The northeastern part of the country has long lasting wet conditions than inland through the hot season. Hence it is expected that the stations from these two contrary characteristics should be presented in distinct pattern groups or clusters. The negative values in silhouette diagram have the meaning of the poor separation in both schemes established for the 6 and 10 cluster levels (Figure 4 and 6). In k-means method, the number of cluster is adjusted by the researcher to get a finer resolution so that the higher level was chosen to get a different viewpoint about the data. But the density of stations in cluster 5 (green colored) remained unchanged (Figure 6).

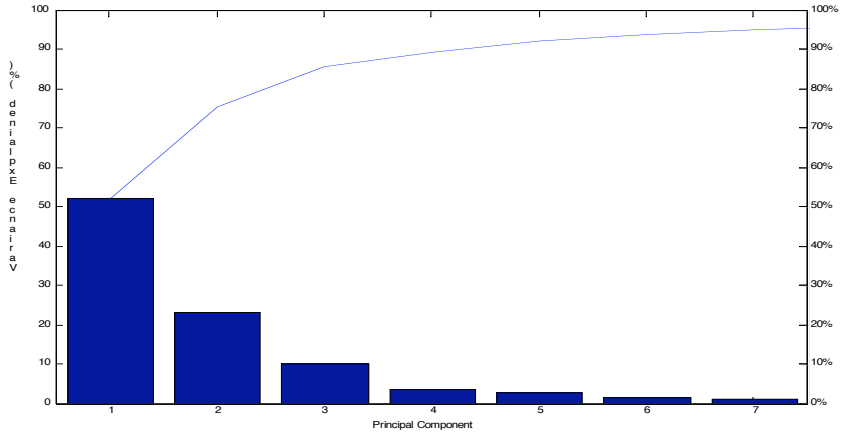

Figure 2. The explained variance percentage by first 7 PCs.

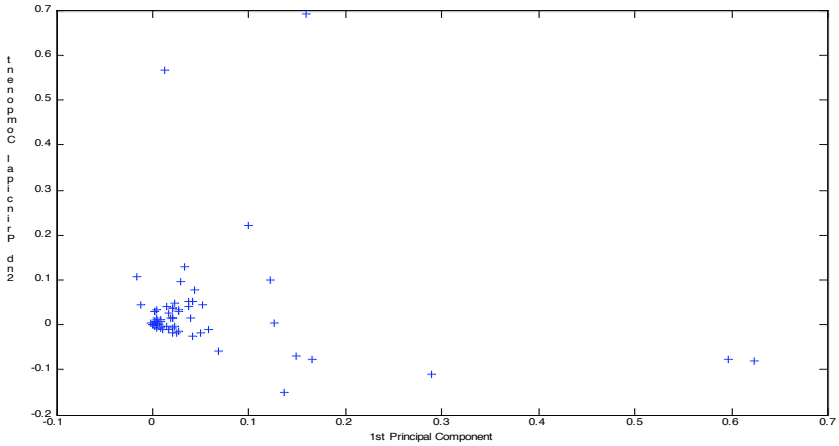

Figure 3. The plot of dispersion of the first two PCs.

The resultant thematic map in Figure 5 has many individual stations in the western part and in the midsection of the country that behaved different from the other regions. This can be explained by one constraint in our data set that is uneven distribution of representative station numbers for each river basin. While the Maritza and Small Menderes basins were represented by only one station, the Sakarya basin was represented by 11 stations in the same spatial resolution. A decrease in the number of stations by gridding watersheds to provide an equal station density can be a solution to avoid this adverse penetration.

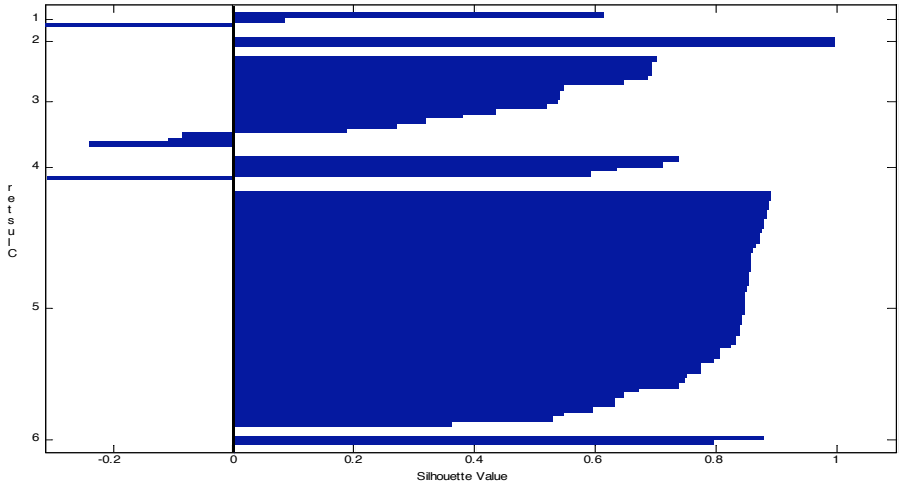

Figure 4. Silhouette diagram for 6 clusters.

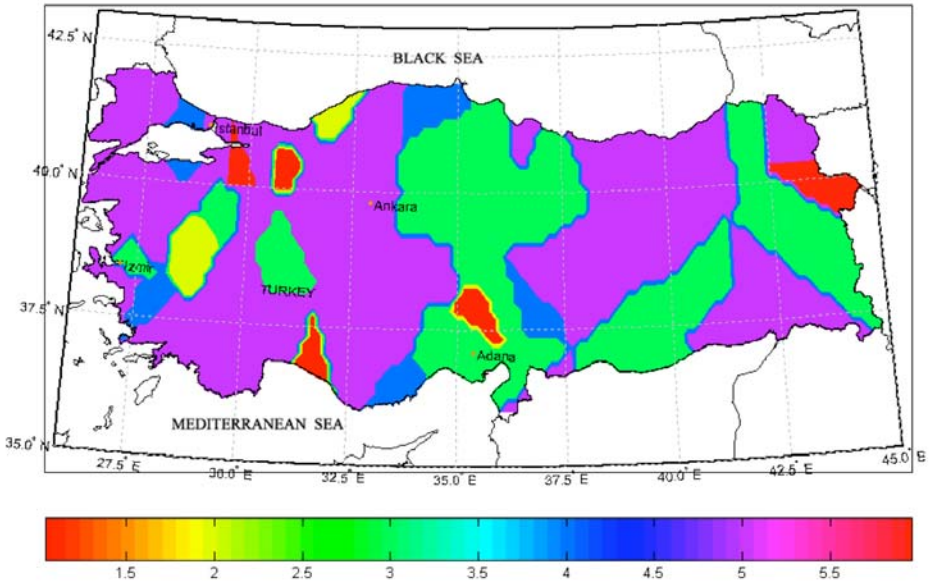

Figure 5. K-means analysis of two principal component scores (Clustering level: **6**)

10 Clustering Level Solution:

In the first part, 6-clustering level was chosen based on the extensive clustering study on the mean monthly streamflow data over Turkey accomplished by Demirel (2004).

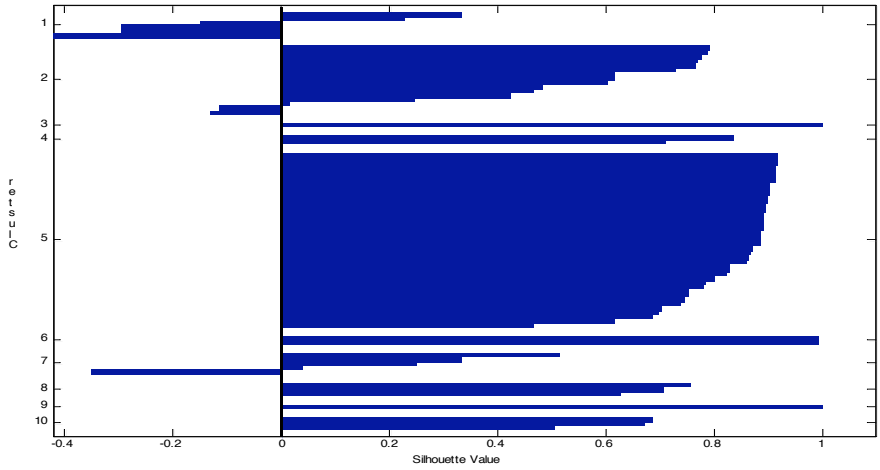

Figure 6. Silhouette diagram for 10 clusters.

In the following diagram, 10-clustering level was mapped. However the change in the cluster density is negligible when the number of level increased. The explanation of a change in streamflow behaviors is often not a simple task, which requires particular analysis at the catchment scale. This is out of the scope of this paper.

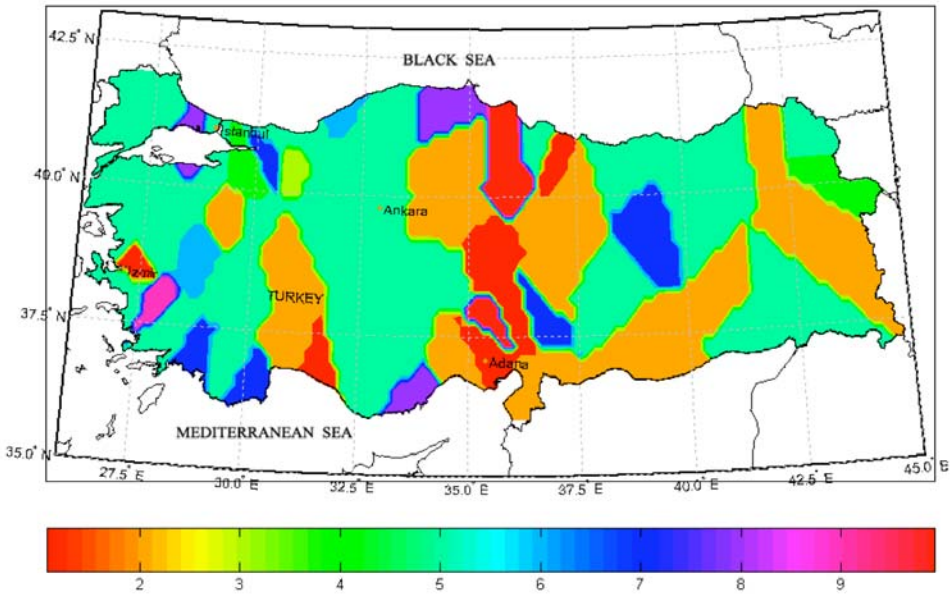

Figure 7. K-means analysis of two principal component scores (Clustering level: **10**)

6. Summary

In this paper, a previous study on monthly minimum streamflow data by Kahya and Demirel (2007) has been extended. Specifically the performance of

PCA and k-means method on meaningful clustering has been scrutinized. The increase in the number of clusters from 6 to 10 did not affect the high density of stations only in one cluster, which is an important indicator of loss of some relevant flow information in our data. The scatter plot of the first two principal component scores showed the same significant unified structure before we applied the method of k-means to the 80x2 data matrix.

Application of PCA on relatively small data set (80x31) is not recommended to use prior to k-means analysis. Further work is needed and is under way to apply cluster analysis in validating short-term intermittent flow prediction models.

Acknowledgements

This research is supported by Istanbul Technical University Research Activities Secretariat (PN# 30695).

References

Andrade E. M. de., 1997: Regionalization of average annual runoff models for ungaged watersheds in arid and semiarid regions. *Ph.D. Thesis*, University of Arizona, Arizona.

Bacher J., 2002. *Cluster Analysis, Lecture Notes*, Nuremberg.

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

Ehrendorfer M.,1987: A regionalization of Austria's Precipitation Climate Using Principal Component Analysis. *Int. J. Climatol.*, **7**, 71-89.

Hisdal H., K. Stahl, L. M. Tallaksen, and S. Demuth, 2001: Have streamflow droughts in Europe become more severe or frequent? *Int. J. Climatol.* **21**, 317–333.

Kahya E., and M. C. Demirel**,** 2007: A Comparison of low-flow clustering methods: streamflow grouping. *Journal of Engineering and Applied Sciences* 2(3): 524-530.

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.

Krasovskaia I., and L. Gottschalk, 1995: Analysis of regional drought characteristics with empirical orthogonal functions. In: *New uncertainty concepts in hydrology and water resources* (ed. by Z.W.Kundezewicz), International hydrology series, Cambridge University press, 163-167.

MacQueen, J. B., 1967: Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, **1**, 281-297.

Nathan R. J., and T. A. McMahon, 1990: Identification of homogeneous regions for the purposes of regionalization. *J. Hydrol.* **121**, 217–238.

Smakhtin V. U., 2001: Low flow hydrology: a review. *J. Hydrol., 240, 147-186.*

Stahl, K., 2001. Hydrological Drought: a study across Europe. *Ph.D. Thesis*, Albert-Ludwigs-Universität, Freiburg.

Stahl K., and S. Demuth, 1999: Linking streamflow drought to the occurrence of atmospheric circulation patterns. Hydrol. Sci. J. **44**(3), 467–482.

Url-1<*http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/kmeans.html*>, accessed at 10.01.2007.

Url-2 <http://www.mathworks.com/>, accessed at 19.01.2007.