Quantification and Reduction of Predictive Uncertainty for Sustainable Water Resources Management (Proceedings of Symposium HS2004 at IUGG2007, Perugia, July 2007). IAHS Publ. 313, 2007.

Assessment of the watershed yield of the Sakarya River basin, Turkey

SABAHATTIN ISIK[1,2] & VIJAY P. SINGH[2]

1 *Department of Civil Engineering, Sakarya University, 54187 Sakarya, Turkey*
isik@tamu.edu

2 *Biological and Agriculture Engineering, Texas A&M University, College Station, Texas 77843, USA*

Abstract The purpose of this study is to classify watershed yields into homogeneous regions and identify the regional membership of watersheds. Monthly river yields of 118 gauging stations in the Sakarya River basin, which is located in northwestern Turkey, were classified by cluster analysis on the basis of hydrological homogeneity. An agglomerative hierarchical clustering algorithm is used so that stations from different geographical locations are considered in the same cluster independently of their geographical location. The study uses Ward's minimum variance linkage method together with the Euclidean distance similarity measures in order to determine the number of homogeneous regions. The Sakarya River basin is clustered into three homogeneous regions and the yield distribution map of the basin is obtained. Out of 118 stations used in this study, 74 gauging stations were grouped into cluster 1, 32 into cluster 2, and 12 into cluster 3. The average watershed yields for each cluster are 0.001966, 0.007336 and 0.01826, respectively, and the average watershed yield of all the basins is 0.005066. Correlation between modelled specific yield and observed yield from gauging stations within each cluster varied from 0.77 to 0.99.

Key words watershed yields; classification; cluster analysis; Sakarya River

INTRODUCTION

Watershed assessments can be used to meet a wide variety of goals. Watershed yield is an important parameter for the planning and design of water structures. In recent years in Turkey it has been emphasized that the development of small hydropower plant projects are needed. The Sakarya River basin is one of most suitable regions for the development of hydropower schemes. However, there are not enough river gauging stations, especially in the desired region or at points where hydropower plants are to be established. It is costly to obtain hydrological information by setting up gauge stations for every desired point. Clustering of subbasins based on hydrometerological homogeneity can be a handy tool in addressing such issues. The classification of gauged watersheds into regions according to preset criteria provides a way to extend information from gauged watersheds to ungauged ones. The purpose of this study is to classify watershed yields into regions and identify the regional membership of gauging stations. Monthly river yields of 118 gauging stations in the Sakarya River basin, which is located in northwestern Turkey, were classified by cluster analysis on the basis of hydrological homogeneity.

Hierarchical cluster analysis techniques are commonly used for hydrological regionalization (Mosley, 1981; Acreman & Sinclair, 1986; Yu & Yang, 1996; Burn *et* *al*., 1997; Stahl & Demuth, 1999; Lecce, 2000; Demirel, 2004; Yanik, 2004). Mosley (1981) applied cluster analysis to data related to the flood hydrology of selected New Zealand catchments and identified regions having similar hydrological regimes. Acreman & Sinclair (1986) classified basins using a multivariate clustering algorithm and tested the homogeneity of basin classes. Yu & Yang (1996) constructed regional flow duration curves for southern Taiwan. Burn *et al*. (1997) reported an approach for catchment regionalization in which an agglomerative hierarchical clustering algorithm was used to define homogeneous regions which can be used for regional flood frequency analysis. Stahl & Demuth (1999) applied a statistical classification method (cluster analysis) to historical streamflow drought time-series and investigated regional patterns of droughts in terms of simultaneous occurrence of streamflow drought at the gauging stations in Europe. Lecce (2000) investigated spatial variations in the timing of the annual flood by using cluster analysis of data from 806 US Geological Survey gauging stations in the southeastern USA. Demirel (2004) applied hierarchical cluster analysis to objectively classify streamflow data from Turkey into regions exhibiting similar streamflow patterns. Yanik (2004) developed regional duration curves for Eastern Black Sea region of Turkey using cluster analysis.

CLUSTER ANALYSIS

Cluster analysis is a multivariate classification procedure that detects natural groupings in data. The goal of a cluster analysis is to group the variables of a data set in a way that the characteristics of the variables within a group are as homogeneous as possible, but the characteristics of the variables between groups are as contrasting as possible. The first and important step is to define a measure (distance) for the similarity or dissimilarity of the characteristics of two variables (Stahl & Demuth, 1999).

The study uses hierarchical cluster analysis, because the number of clusters is not known *a priori*. Hierarchical clustering techniques can be classified into agglomerative methods and divisive methods. The agglomerative method proceeds by a series of successive fusions of n individuals into groups, but divisive methods proceed by separating the n individuals successively into finer groupings. Both types of hierarchical clustering attempt to find the most efficient step at each stage in the progressive subdivision of data. The hierarchical clustering method is best represented by a two-dimensional diagram, known as a dendrogram or tree diagram, which illustrates the divisions made at each successive stage of analysis (Everitt *et al*., 2001).

The cluster number is found by using Ward's approach which is an agglomerative hierarchical cluster analysis method. The Ward method seeks to form 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. At each step in the analysis, the union of every possible cluster pair is considered and two clusters whose fusion results in a minimum increase in the information loss are combined. Information loss is defined by the Ward method in terms of the error sum-of-squares criterion. The Euclidean distance function is used as a similarity metric to measure the distance d_{ij} between two objects i and j. The Ward method tries to minimize the total within-group or within-cluster sums of squares as a measure of homogeneity (Rencher, 2002).

In order to estimate streamflow at ungauged sites, Fennesy & Vogel (1990), Chiang *et al*. (2002), Murdock & Gulliver (1993), among others, developed the regression:

$$Q_u = qA_u \quad (1)$$

where Q_u is the estimated streamflow at an ungauged site; A_u is the watershed area at the ungauged site; and q is the specific stream flow or specific water yield of a watershed defined as:

$$q = \frac{Q}{A} \quad (2)$$

where Q is the stream flow discharge ($m^3 s^{-1}$) and A is the drainage area (km^2). Specific streamflow is the discharge divided by the basin area and has the units: $m^3 s^{-1} km^{-2}$. Monthly streamflow values, calculated by averaging daily streamflow within a month, and drainage area are important parameters for hydrological regionalization.

STUDY AREA AND DATA

The Sakarya River is situated in the northwest Anatolian region of Turkey and its length is nearly 810 km. The drainage area of the Sakarya River, which equals approximately 1/13 of the total area of Turkey, is about 56 000 km^2. The regionalization method was applied to monthly discharges of 118 river gauging stations in the Sakarya River basin (Fig. 1). All river gauging stations with at least five years of measurements were taken into account. Monthly average watershed yields or specific streamflow discharge ($m^3 s^{-1} km^{-2}$) was calculated by dividing monthly streamflow discharges ($m^3 s^{-1}$) by the drainage area (km^2) and was normalized between the range of 0 and 1. Streamflow measurements in Turkey are carried out by the State Hydraulic Works (DSI) and Electrical Power Resources Survey and Development Administrations (EIE). The mean and standard deviation of monthly average yield are given in Table 1 (DSI, 2000; EIE, 2003). The mean value of monthly average yield for all station varies between 0.00132 and 0.01115, and standard deviation values vary between 0.00146 and 0.01281 for 12 months.

Table 1 Statistical parameters of data ($m^3 s^{-1} km^{-2}$).

Months	Mean	Std deviation
1	0.00187	0.00264
2	0.00282	0.00495
3	0.00498	0.00747
4	0.00617	0.00776
5	0.00787	0.00852
6	0.01087	0.01072
7	0.01115	0.01281
8	0.00678	0.00867
9	0.00365	0.00458
10	0.00192	0.00229
11	0.00139	0.00172
12	0.00132	0.00146

Fig. 1 Sakarya River basin.

The region has a mild climate, characterized by rain throughout the year with heaviest rains occurring in winter and early spring months. Rainfall averages about 770 mm per year and is fairly well distributed throughout the year. Snowfall is relatively light with an average between 0.2–0.3 m depth per year.

In the clustering process, standardization is essential when variances among variables differ to any great extent; therefore, standardization techniques were applied to data. However, the original data were used in the analysis to compare the results of magnitude differences among stations.

RESULTS

The Euclidean distance and the Ward method of agglomerative hierarchical clustering were used to determine homogenous regions. The tools from MATLAB were used for clustering. In order to interpret the dendrogram of clustered data, distances were read from the dendrogram and plotted against the number of clusters. The dendrograms of clustered data and distance test (DAN) are shown in Figs 2 and 3.

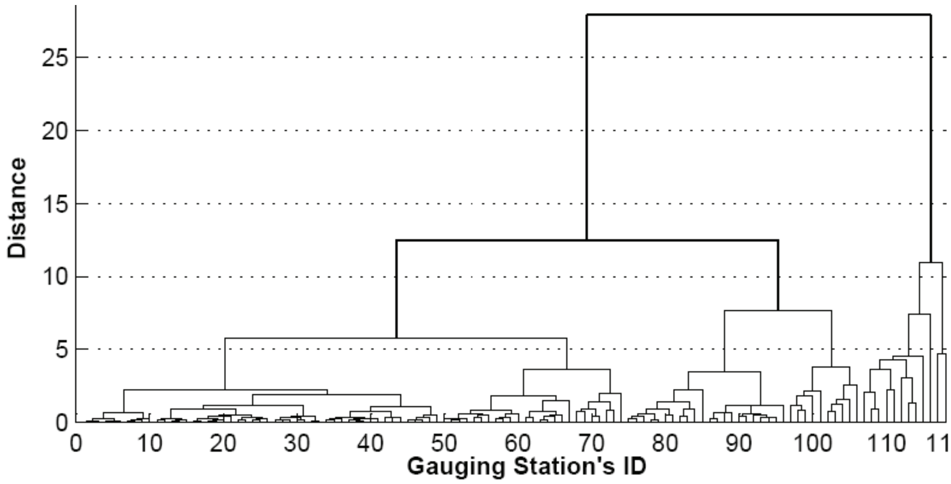

Fig. 2 Dendrogram of 118 river gauging stations.

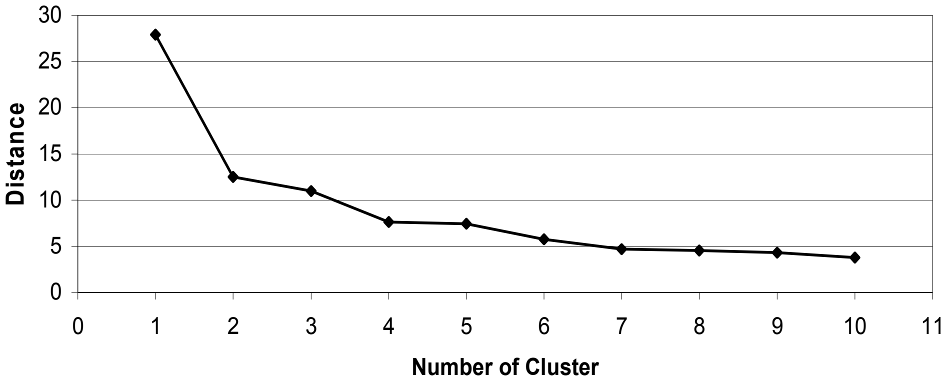

Fig. 3 Distance test.

The RMSSTD and RSQ values for each cluster are shown in Fig. 4(a) and (b), respectively. It is apparent from Figs 2 and 3 that the total of 118 gauging stations may be grouped into four clusters. However, the last two stations of the right hand side in the dendrogram assemble in Cluster 4. Since the two stations are far from each other, they do not compose a group. Either they are allocated to the closest homogenous group or discarded if they contribute greatly to the heterogeneity in the homogenous region (Burn, *et al.*, 1997). If the dendrogram is cut off at 3 as seen from Fig. 2, the last two stations of the right hand side are allocated to Cluster 3. Therefore, average monthly specific discharges (m^3 s^{-1} km^{-2}) of the 118 river gauging stations in Sakarya

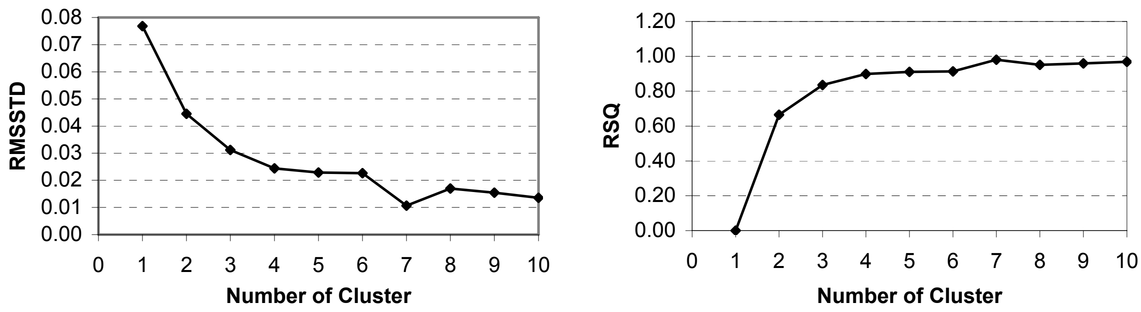

Fig. 4 (a) RMSSTD test. (b) RSQ test.

Table 2 Statistical characteristics of homogenous regions.

Months	Cluster 1 (74 stations)		Cluster 2 (32 stations)		Cluster 3 (12 stations)	
	Mean	Std dev.	Mean	Std dev.	Mean	Std dev.
1	0.000938	0.000596	0.001836	0.001329	0.007716	0.004883
2	0.001078	0.000635	0.003039	0.00237	0.013002	0.010434
3	0.001502	0.000859	0.007012	0.00387	0.020992	0.013195
4	0.00221	0.001225	0.008467	0.003925	0.024509	0.009952
5	0.003135	0.001869	0.011357	0.004776	0.027791	0.008043
6	0.004036	0.002473	0.018505	0.00519	0.032611	0.010332
7	0.003573	0.002014	0.019552	0.010131	0.035452	0.013938
8	0.002541	0.001292	0.010055	0.005381	0.024211	0.014926
9	0.001685	0.001178	0.00449	0.002793	0.013494	0.007602
10	0.001095	0.001028	0.001725	0.001394	0.007509	0.002162
11	0.00091	0.000931	0.001028	0.000933	0.005284	0.002178
12	0.000891	0.00073	0.000964	0.000741	0.004947	0.001301

River basin were divided into three homogenous regions. Statistical characteristics of homogenous regions are given in Table 2. The average specific streamflows are 0.001966, 0.007336, and 0.018126 for Clusters 1, 2 and 3, respectively; while the average watersheds yield of the basin is 0.005066. Homogenous regions in the Sakarya River basin are shown in Fig. 5.

The first cluster has 74 stations in which average monthly specific streamflows vary between 0.000891 and 0.004036. Correlation coefficients between modelled specific yield and observed specific yield are calculated for each station. Correlation coefficients between modelled values and observed specific streamflows of 74 stations within Cluster 1 vary from 77% to 99.7%. It has the lowest specific streamflows of the three clusters. Cluster 1, which has the largest area and is in the upstream part of the Sakarya basin, is a semi-arid region. Most of the hydro structures, such as hydro-electric, flood control, irrigation, and domestic purposes dams, are located in Cluster 1.

The second cluster has 32 stations in which average monthly specific streamflows and standard deviation vary 0.000964–0.019552 and 000741–0.01031, respectively. Correlation coefficients between modelled value and observed specific streamflows of 32 stations within Cluster 2 vary from 92% to 99.7%. Cluster 2 is located in the northern part of the Sakarya basin where the precipitation ratio is higher than Cluster 1. It is not a semi-arid region like Cluster 1.

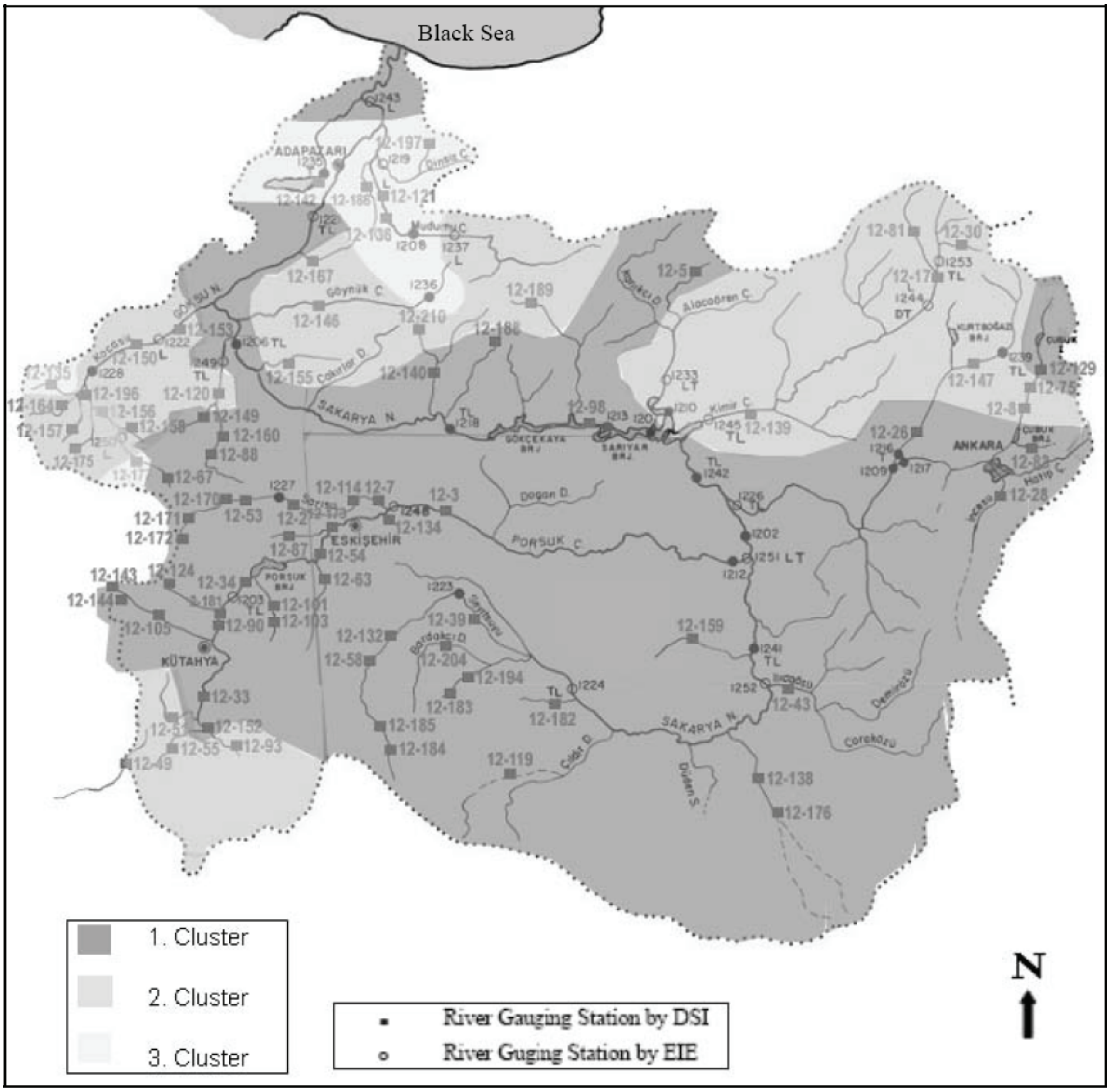

Fig. 5 Homogenous regions in the Sakarya River basin.

The third cluster has 12 stations and the highest average monthly specific discharges; average monthly specific streamflows and standard deviation range 0.004947–0.035452 and 001301–0.014926, respectively. Correlation coefficients between modelled value and observed specific streamflows of the 12 stations within Cluster 3 vary from 98% to 99%. The third cluster is mostly alluvial in the downstream part of the Sakarya basin. The participation ratio is the highest and this region has frequent floods. While the regions of Cluster 1 and 2 are generally forest areas with steep topography, Cluster 3 is flat and has a plain topography.

It can be concluded that each homogeneous cluster is proportional to its physiographic characteristics, climate, and land-use pattern in the Sakarya basin. As the area of Cluster 1 is largest, it can be clustered to obtain sub-homogeneous regions for Cluster 1 for further studies.

CONCLUSION

Monthly average yields ($m^3 s^{-1} km^{-2}$) of 118 river gauging stations in the Sakarya River basin were classified into three homogenous regions. An agglomerative hierarchical clustering algorithm was used to cluster data. In order to determine the number of homogeneous regions, the root mean square standard deviation (RMSSTD), *r*-squared (RSQ), and Distance Analysis (DAN) test were conducted. Out of 118 stations used in this study, 74 gauging stations were grouped into Cluster 1, 32 into Cluster 2 and 12 into Cluster 3, and the average watershed yield of all the basins was 0.005066. The correlation coefficient between observed and modelled water yields from each gauging station within a cluster varies from 0.77 to 0.99. It is concluded that each homogeneous cluster is related to its physiographic characteristics, climate, and land-use pattern in the Sakarya basin.

REFERENCES

Acreman, M. C. & Sinclair, C. D. (1986) Classification of drainage basins according to their physical characteristics: an application for flood frequency analysis in Scotland. *J. Hydrol.* **84**(3-4), 365–380.

Burn, D. H., Zrinji, Z. & Kowalchulk, M. (1997) Regionalization of catchments for regional flood frequency analysis. *J. Hydrol. Engng* **2**(2), 76–82.

Chiang, S. M., Tsay, T. K. & Nix, S. J. (2002) Hydrologic regionalization of watersheds. I: Methodology development. *J. Water Resour. Plann. Manage. ASCE* **128**(1), 3–11.

Demirel, M. C. (2004) Cluster analysis of flow data over Turkey. MSc Thesis, Istanbul Technical University, Istanbul, Turkey.

DSI (2000) Monthly mean flows of Turkish rivers. Ankara, Turkey.

EIE (2003) Monthly mean flows of Turkish rivers, (1935–2000). Ankara, Turkey.

Everitt, B. S., Landau, S. &Leese, M. (2001) *Cluster Analysis.* Oxford University Press, London, UK.

Fennessey, N. & Vogel, R. M. (1990) Regional flow duration curves for ungaged sites in Massachusetts. *J. Water Resour. Plann. Manage. ASCE* **116**(4), 530–549.

Lecce, S. A. (2000) Spatial variations in the timing of annual floods in the Southeastern United States. *J. Hydrol.* **235**, 151–169.

Mosley, M. P. (1981) Delimitation of New Zealand hydrologic regions. *J. Hydrol.* **49**, 173–192.

Murdock, R. U. & Gulliver, J. S. (1993) Prediction of river discharges at ungauged sites with analysis of uncertainty. *J. Water Resour. Plann. Manage. ASCE* **119**(4), 473–487.

Rencher, A. C. (2002) *Methods of Multivariate Analysis.* John Wiley & Sons Inc., New York, USA.

Stahl, K. & Demuth, S. (1999) Methods of regional classification of stream flow drought series: cluster analysis. *ARIDE, Technical Report no. 1*. Institute of Hydrology, University of Freiburg, Germany.

Yanik, B. (2004) Dogal akisli hidroelektrik potansiyelin belirlenmesinde bolgesel analiz yaklasimi. PhD Dissertation, Istanbul Technical University, Istanbul, Turkey.

Yu, P. S. & Yang, T. C., (1996) Synthetic regional flow duration curve for Southern Taiwan. *Hydrol. Processes* **10**, 373–308.