

Some problems in regionalization of watersheds

A. RAMACHANDRA RAO

School of Civil Engineering, Purdue University, West Lafayette, Indiana 47907, USA
rao@ecn.purdue.edu

V. V. SRINIVAS

Department of Civil Engineering, Indian Institute of Science, Bangalore 560012, India

Abstract Because of the paucity of flood data, it is not always possible to use methods of frequency analysis to estimate flood values corresponding to specified recurrence intervals. To alleviate this problem, hydrologists have used information from watersheds in a region to estimate the magnitude of floods. Several regionalization procedures, which attempt to identify watersheds with a similar hydrological response, have been developed in the literature. However, as no single procedure has been demonstrated to yield universally acceptable results, several methods of regionalization are in use. In this study, three hybrid-cluster algorithms, which are a blend of agglomerative hierarchical and partitional clustering procedures, are tested to determine their potential in combining watersheds in Indiana, USA, into homogeneous regions. The problems encountered in this investigation are discussed.

Key words clustering algorithms; flood flows; regionalization

INTRODUCTION

In hydrology, approaches which have been used for regionalization of watersheds include: (a) the method of residuals; (b) the canonical correlation analysis (Cavadias *et al.*, 2001); (c) the region-of-influence (ROI) approach (Burn, 1989) and its extensions; (d) the hierarchical approach (Gabriel & Arnell, 1991) and its extension to the ROI framework; and (e) cluster analysis (Burn, 1989; Bhaskar & O'Connor, 1989; Burn & Goel, 2000). As none of these procedures has been demonstrated to yield universally acceptable results, several methods of regionalization are in use. No information is available about the relative performance of these methods. Some problems encountered in regionalization of watersheds with hard clustering methods are discussed in this paper. In particular, problems with the use of Hybrid Cluster Analysis in regionalization, are discussed.

Hard clustering algorithms can be broadly classified into two categories: hierarchical and partitional clustering. Hierarchical clustering algorithms fall into two categories: agglomerative and divisive. The partitional clustering procedures require an initial guess about the number of clusters and cluster centres. They can be classified by using the technique used to initiate clusters, clustering criteria and the type of data for which they are applicable. The K-means algorithm and agglomerative hierarchical clustering algorithms have been used for regionalization in hydrology (Mosley, 1981; Tasker, 1982; Nathan & McMahon, 1990; Burn, 1989).

In this study, three hybrid-cluster algorithms, which are a blend of agglomerative hierarchical and partitional clustering procedures, were used to determine their potential

in delineating watersheds in Indiana, USA, into regions that are homogeneous in hydrological response. The hierarchical clustering algorithms considered for hybridization were single linkage, complete linkage and Ward's algorithms, while the partitioning clustering algorithm used is the hard K-means algorithm.

After the regions were established by using the clustering algorithms, their homogeneity was tested using the statistics of Hosking & Wallis (1993, 1997). Three heterogeneity measures H_1 , H_2 and H_3 of dispersion suggested by Hosking & Wallis (1993) were used for the analysis. A region can be regarded as "acceptably homogeneous" if $HM < 1$, "possibly homogeneous" if $1 \leq HM < 2$, and "definitely heterogeneous" if $HM \geq 2$, where HM is the heterogeneity measure. Further details on the homogeneity test are found in Hosking & Wallis (1997). These clusters were then revised to make the regions more homogeneous. The revision often required exclusion of those sites from a cluster that are grossly discordant with respect to other sites of the cluster.

DATA USED IN THE STUDY

Flow records from 273 gauging stations in Indiana were used in the study. These included all the data from 245 stations considered by Glatfelter (1984) (Fig. 1). In pooling data from additional stations to those considered by Glatfelter, a screening criterion of having a minimum record length of 10 years was used. Twenty-eight stations passed this screening test. The selection of the record length threshold for the screening process is subjective. In the discussion to follow, the 28 stations included through the screening criterion will be referred to as "pooled" stations. The location of these stations in the study region is shown in Fig. 2.

Sensitivity of flood response of drainage basins to variation in the values of any attribute is examined by plotting each one of the attributes against a flood-related variable. The objective of this exercise was to identify independent attributes. The flood related variables considered in this study included the mean and median values of annual flood, mean and median values of annual flood per unit area of drainage basin, mean annual flood divided by the mean annual precipitation, median annual

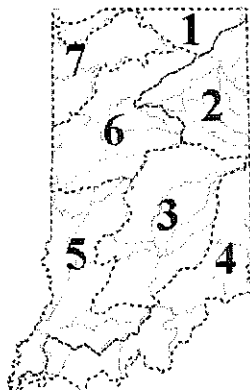


Fig. 1 The seven hydrological regions identified by Glatfelter (1984) for estimating the magnitude and frequency of floods on streams in Indiana.

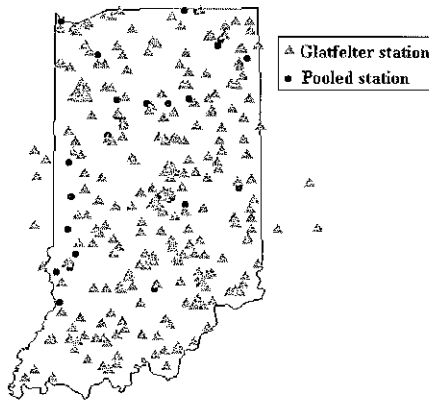


Fig. 2 Geographic location of streamflow gauging stations considered for regionalization of the study region.

flood divided by the mean annual precipitation (Table 1). Finally, the features extracted for cluster analysis are: (a) four physiographic attributes (drainage area, slope of the main channel in the drainage basin, soil runoff coefficient and storage); and (b) one meteorological attribute (mean annual precipitation). The latitude and longitude were included in the feature vector with a view to identify regions that are geographically contiguous. Information pertaining to all these attributes was available for the 245 stations considered by Glatfelter (1984). Data concerning the slope of the main channel in the drainage basin, soil runoff coefficient and percentage of drainage area covered by lakes, ponds or wetlands were not available for the 28 pooled stations. Therefore, only the 245 gauging stations considered by Glatfelter (1984) were used in cluster analysis and the 28 pooled stations were included in the resulting clusters which contain them geographically. When the entire set of 273 sites was considered as a single cluster, the region was highly heterogeneous ($H_1 = 15.73$, $H_2 = 5.81$ and $H_3 = 1.78$). This justifies the need to identify homogeneous groups of watersheds in Indiana. Of the seven attributes only drainage area was transformed using logarithmic transformation. Each of the seven attributes was standardized such that their mean is zero and variance is unity.

Table 1 Attributes available for the Glatfelter stations.

Attribute	Range
Drainage area	0.11–11125.00 mile ²
Mean annual precipitation	34–46 in
Main channel slope	0.90–267.00 ft per mile
Main channel length	0.3–315.0 mile
Basin elevation	412.0–1190.0 ft
Storage	0–11%
Soil runoff coefficient	0.30–1.00
Forest cover in drainage area	0.0–88.4%
$I(24,2)$	2.6–3.35 in

$I(24,2)$: 24-h rainfall having a recurrence interval of 2 years, in inches.

Storage: percentage of the contributing drainage area covered by lakes, ponds or wetlands.

RESULTS AND DISCUSSION

The K clusters obtained from the agglomerative hierarchical clustering constituent of the hybrid model after $245-K$ merges were used to initiate the K-means algorithm. The objective function, in general, decreases with increase in the number of clusters. It has maximum value when all the feature vectors are lumped into a single cluster and has a minimum value of zero when K equals the number of feature vectors considered for cluster analysis.

Variation in the optimal value of the objective function for K ranging from 1 to 10 is presented in Fig. 3 for each of the three hybrid clustering models and their respective hierarchical clustering constituents, namely single linkage (SL), complete linkage (CL) and Ward's hierarchical clustering algorithms. Of the three hierarchical clustering constituents of the hybrid model, Ward's algorithm gave the smallest value for the objective function. As expected, the performance of the hybrid clustering algorithms is better than their respective hierarchical clustering constituents. In particular, the blend of Ward's and hard K-means algorithms has the minimum values of the objective function for most values of K in the range from 1 to 10. The blend of complete linkage and hard K-means algorithms yielded the smallest values of objective function for the choice of K equal to 4 and 6.

The clusters obtained from the three hybrid clustering algorithms were also examined by plotting them on a map of the state. Despite some differences in the value of the objective function, the clusters resulting from the hybrid clustering algorithms were not very different from one another. The clusters obtained from the hybrid of Ward's algorithm and K-means algorithm were found to be very similar to those resulting from Ward's algorithm., which implies a small role played by the K-means

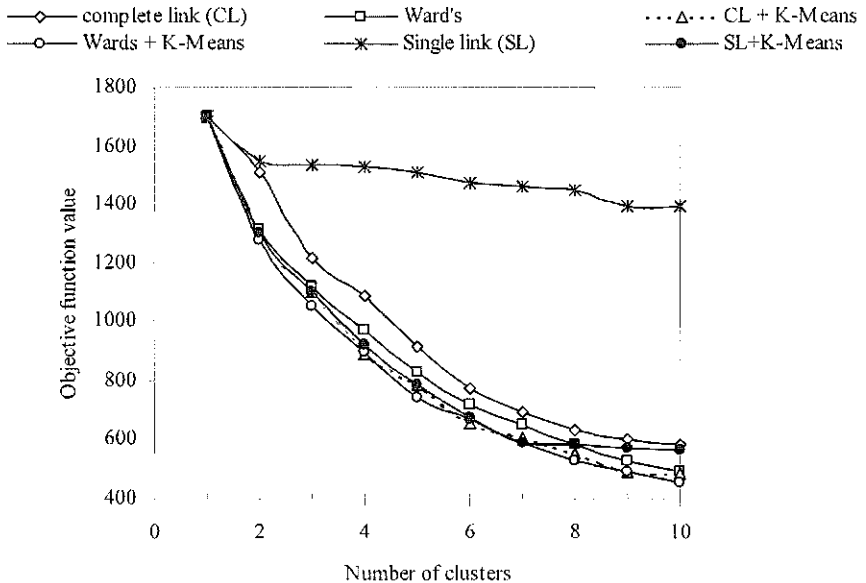


Fig. 3 Variation of objective function with increase in the number of clusters: comparison of the hybrid models with their hierarchical clustering constituents.

algorithm in arriving at the final clusters. In contrast, the clusters resulting from the single linkage algorithm were considerably modified by the K-means algorithm.

The clusters obtained from the combination of Ward's algorithm and the K-means algorithm for different choices of K are compared in terms of their size (i.e. number of sites and their collective record length) and heterogeneity indices. The group of clusters, corresponding to a choice of K , for which a major part of the feature vectors considered for analysis are found in clusters that have the values of heterogeneity measures (H_1 , H_2 and H_3) closer to being homogeneous, are identified. These clusters are then revised to improve their homogeneity measures and the physical coherence of the regions.

Increase in the number of clusters resulted in segregation of a collection of sites that are highly heterogeneous. This cluster comprises of drainage basins located in the Kankakee basin in the northwestern part of Indiana. The plausible regions resulting from the clustering algorithm for the choice of K equal 10 were quite satisfactory and were analysed further. A significant number of basins in cluster 2 have small lakes. Mild slopes and low runoff coefficient values characterize the catchments in this cluster. This cluster consisting of five stations was identified when the value of K is increased beyond 4. Cluster 3, consisting of 11 stations, is located in the karst region of southern Indiana and it comprises of sites that have steep slopes and high soil runoff coefficient values.

The groups of stations resulting from cluster analysis were revised to obtain hydrologically homogeneous regions. The options that were considered for revising the plausible regions resulting from a clustering algorithm are: (a) excluding one or more sites from a region to join a residual set; (b) transferring one or more sites from a region to other regions; (c) dividing a region to form two or more new regions; (d) allowing one or more sites near the border between two regions to belong to both the regions; (e) dissolving a region by transferring its sites to other regions; (f) merging two or more regions. Of these, the first four options are useful in reducing the values of heterogeneity measures, whereas the last two options were of help in ensuring that each region is sufficiently large. The effort required for the task of merging or splitting a region is minimal when a clustering algorithm is used to obtain regions, because it provides a variety of plausible scenarios of regions to increase the number of clusters.

Of the ten clusters, region 1 resulted in revising the collection of these stations. The second cluster had just five sites. This cluster was subsequently dissolved by transferring these sites to other regions. Region 3 resulted from revising the fourth cluster that characterizes drainage basins with a high value of soil runoff coefficient. Cluster 3 has just 133 data points making it the smallest of all clusters in terms of information. It is merged with cluster 9 that contained it geographically and revised to obtain region 2. Region 4 resulted when the first and seventh clusters were merged and revised. Clusters 1 and 7 consist of geographically neighbouring drainage basins that have similar soil runoff characteristics, mean annual precipitation and surface storage features. They are, however, quite distinct in their drainage areas and slope of main streams draining the basins. Cluster 1 consisted of basins with steep slopes and smaller drainage areas, while cluster 7 contained basins with mild slopes and larger drainage areas. Basins in clusters 1 and 7 appear as a single group for a choice of K equal to 4, 5 or 6. Hence these clusters can be merged.

Table 2 Results from the exercise performed to test the regions for robustness.

Region number	Condition	Number of stations	Heterogeneity measure:			Region type
			H_1	H_2	H_3	
1	Entire region	59	0.86	-0.12	-0.94	Homogeneous
	Sites with RL ≤ 10 eliminated	41	1.08	0.07	-0.73	Possibly homogeneous
	Sites with RL < 20 eliminated	31	0.88	0.40	-0.24	Homogeneous
	Sites with RL ≥ 35 eliminated	51	0.39	-0.24	-0.75	Homogeneous
2	Entire region	58	0.85	0.43	-0.65	Homogeneous
	Sites with RL < 20 eliminated	36	0.58	1.16	0.56	Homogeneous
	Sites with RL > 50 eliminated	49	0.88	0.48	-0.76	Homogeneous
	Sites with RL ≤ 10 and RL > 50 eliminated	40	0.85	0.87	-0.17	Homogeneous
3	Entire region	30	-0.46	0.66	0.28	Homogeneous
	Sites with RL ≤ 10 eliminated	21	-0.28	0.40	0.20	Homogeneous
	Sites with RL < 20 eliminated	15	-0.41	0.01	-0.10	Homogeneous
4	Entire region	73	0.48	-0.40	-1.78	Homogeneous
	Sites with RL < 20 eliminated	63	0.81	0.12	-1.03	Homogeneous

RL: record length.

The heterogeneity measures of Hosking & Wallis (1993) weigh information from each station in proportion to its record length. As a consequence, the influence of stations with a longer record length will be higher than those with a shorter record length. This may have adverse effects, especially when some stations in a region have much longer record lengths than others. Therefore, the hydrological regions that were obtained by revising regions were examined for their robustness. In this exercise, specifying various threshold values segregates stations with record lengths significantly different from rest of the group, and the region comprising of the rest of the stations was examined for homogeneity. It was also intended to identify and exclude a few stations that have an adverse affect on the homogeneity of the regions. The results from this exercise are presented in Table 2. It is evident that all the homogeneous regions identified are indeed robust.

The results presented in Table 3 indicate that regions 1 to 5 are acceptably homogeneous, while region 6 adjoining Lake Michigan is highly heterogeneous. Region 1 is spread mainly along the course of the Wabash River and comprises predominantly of alluvial deposits of the flood plains (Fig. 4). Region 2 contains karst formations associated with limestones of the Mississippian age, laid down 320–360 million years ago. Region 3 possesses a karst area consisting of older Devonian and Silurian limestones. The topography of these areas is dominated by sinkholes, sinking streams, large springs and caves. For the ungauged catchments lying at the border between regions 2 and 3, the possibility of including information from both the regions

Table 3 Characteristics of the regions formed.

Region number	N	RS	Heterogeneity measure:		
			H_1	H_2	H_3
1	62	1689	0.86	-0.12	-0.94
2	58	1730	0.85	0.43	-0.65
3	30	804	-0.46	0.66	0.28
4	73	3039	0.48	-0.40	-1.78
5	42	1938	0.04	-0.91	-0.85
6	14	519	13.69	6.33	2.94

N: Number of stations.

RS: Region size in station years.

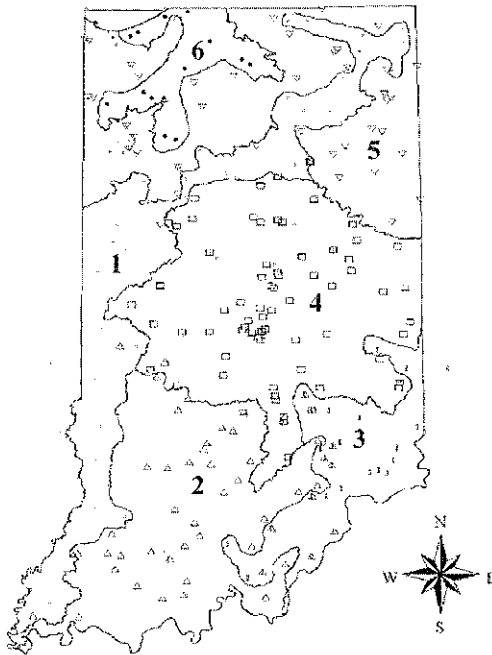


Fig. 4 Location of the regions defined using the hybrid cluster analysis.

can be considered. Region 4 is in central Indiana. The soil here is predominantly loamy glacial till. Region 5 extends over the northern part of Indiana. It comprises of a wide range of soil classes (clayey glacial till, sandy and loamy deposits, loamy glacial till) overlying the Mississippian rocks of the Michigan basin and Devonian and Mississippian shale.

CONCLUSIONS

The following conclusions may be presented on the basis of the results discussed herein:

- (a) The selection of attributes plays a crucial role in arriving at groups of watersheds which may be similar in response.
- (b) The combination of Ward's and the K-means algorithms give good initial estimates of groups of watersheds.
- (c) The homogeneity measures must be used to decide the group of watersheds which are to be included in a cluster.
- (d) Testing the homogeneity of selected regions by using homogeneity measures is important.
- (e) Simply using the clustering algorithms will not result in homogeneous regions.

Acknowledgements This paper is a part of the project "Regionalization of Indiana Watersheds for Flood Flow Predictions (Phase I)" supported by the Joint Transportation Research Program (Indiana SPR-2476, 656-1284-0137, file: 9-8-11). We would like to acknowledge the support by JTRP, which made the study possible. Merril Dougherty, David Finley and Dave Ward of INDO'T, Keith Hoernschemeyer of FHWA, David Knipe and Raj Gosine of IDNR served on the Study Advisory Committee. They were also of considerable help in conducting the study. Ms Karen Hatke, Administrator of JTRP was of considerable and cheerful help in administrative matters. Mrs Dinah Hackerd assisted in numerous ways, especially in preparing the report. It is a pleasure to thank them for all their help.

REFERENCES

- Bhaskar, N. R. & O'Connor, C. A. (1989) Comparison of method of residuals and cluster analysis for flood regionalization. *J. Water Resour. Plan. Manage.* **115**(6), 793-808.
- Burn, D. H. (1989) Cluster analysis as applied to regional flood frequency. *J. Water Resour. Plan. Manage.* **115**(5), 567-582.
- Burn, D. H. & Goel, N. K. (2000) The formation of groups for regional flood frequency analysis. *Hydrol. Sci. J.* **45**(1), 97-112.
- Cavadias, G. S., Ouarda, T. B. M. J., Bobee, B. & Girard, C. (2001) A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins. *Hydrol. Sci. J.* **46**(4), 499-512.
- Gabriele, S. & Arnell, N. (1991) A hierarchical approach to regional flood analysis. *Water Resour. Res.* **27**(6), 1281-1289.
- Glatfelter, D. R. (1984) Techniques for estimating magnitude and frequency of floods on streams in Indiana. US Geological Survey, Water Resources Investigations Report 84-4134.
- Hosking, J. R. M. & Wallis, J. R. (1993) Some statistics useful in regional frequency analysis. *Water Resour. Res.* **29**(2), 271-281. Correction: *Water Resour. Res.* **31**(1), p.251, 1995.
- Hosking, J. R. M. & Wallis, J. R. (1997) *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press, UK.
- Mosley, M. P. (1981) Delimitation of New Zealand hydrological regions. *J. Hydrol.* **49**, 173-192.
- Nathan, R. J. & McMahon, T. A. (1990) Identification of homogeneous regions for the purposes of regionalization. *J. Hydrol.* **121**, 217-238.
- Tasker, G. D. (1982) Comparing methods of hydrologic regionalization. *Water Resour. Bull.* **18**(6), 965-970.