DELINEATION OF HOMOGENEOUS FLOOD REGIONS IN CENTRAL INDIA USING SINGLE AND COMPLETE LINKAGE METHODS

BHAGABAT P. PARIDA
Department of Environmental Science, University Of Botswana, Private Bag UB0704, Gaborone, Botswana.

Methods of single and complete linkage were used for clustering the flood sites of central India, using attributes such as maximum rainfall in design duration, average \( \Phi \)-index, catchment area, basin slope, length of main stream, and mean annual peak flood data, which have been considered to be responsible for flood generation. Based on clear-cut clusters on the dendrogram. Complete linkage method was selected as appropriate method for the study area. Homogeneity of identified clusters (regions) and choice of appropriate distributions have been undertaken using method on L-Moments and growth curve for each region developed. Relation between mean annual peak flood data and basin characteristics, have also been developed for each homogeneous region using various catchment characteristics.

INTRODUCTION

The internal transitions of the earth and lack of monotony in energy received from sun are some of the basic reasons for creating heterogeneous geographical regions. On the other hand the similarity in environment and natural events state the homogeneity. From a hydrological perspective, homogeneous regions are usually defined as geographical regions having similar hydrological, climatic and physiographic characteristics. In such cases, information from gauged sites can be transferred to ungauged ones while being able to improve at-site estimation of flood quantiles. Mainly, flood regionalization techniques involve three basic steps: delineation of homogeneous regions, agglomeration of the gauged flood data using a common statistical distribution and development of regional relationship between mean annual flood and catchment characteristics, such that results of regional analyses can be transferred to ungauged locations in the identified region. However, for quantile estimation at either gauged and ungauged locations, while the steps two and three are useful, there seems to be no uniquely advocated objective approach to the delineation of homogeneous regions, though homogeneity for identified regions are usually tested using some statistical indices.

Amongst many approaches, hierarchical clustering approaches have widely found place in various fields of science such as geology, chemistry, fisheries, and medical science and especially in taxonomy of organisms, as compared to their use on flood or related variables. However, a brief review of the application of various forms of hierarchical clustering methods to flood analysis reveals that, Moseley [7] successfully used hierarchical clustering procedure for delineation of homogeneous regions in New
Zealand, using two methods for illustration of clusters viz: illustration through the plot of specific mean annual floods versus coefficient of variation of flood at different sites and through construction of dendrogram. Further, Tasker [9] made use of the above concept together with that of Decoursey [3] but used a complete linkage algorithm to form homogeneous clusters using similarity of hydrologic characteristic in 221 stations of Arizona. In yet another study, Acreman and Sinclair [1] used hierarchical procedure for delineation of peak flood homogeneous regions based on basin characteristics data in Scotland.

Gottschalk [4] used yet another method of hierarchical clustering such as the Unweighted Pair Group Method (UNWPGM) and the method of Principal Component Analysis (PCA) for clustering the flood sites of Sweden. His finding was that, the UNWPGM was useful in nation scale classification whereas the PCA, which is a non-hierarchical technique, was useful for further classification of these broad scale regions. Besides these, even recently Kim and Hawkinings [6] and Zrinji and Burn [10] have demonstrated the uses of the hierarchical approach for determination of homogeneous regions in North-eastern U.S.A and West-central Canada respectively. In view of these, this study aims to make use of hierarchical techniques in two different forms namely the Single Linkage and the Complete Linkage Methods using attributes responsible for flood generation process (Parida, [8]), to delineate flood sites in one of the most flood ravaged zone in Central India and the resulting regions verified for their Homogeneity using L-Moments method and finally for development of regional curves/equations.

**METHODOLOGY**

General procedure for carrying out the agglomerative technique in the hierarchical category can be summarized through the following steps.
1- Data matrix of size (n x p) is constructed using data from n gauged sites each having p number of variables.
2- Determination of similarities or dissimilarities among the sites leading to organization of similarity or dissimilarity matrix. (n x n)
3- Looking for the maximum value within the similarity matrix (or minimum value within the dissimilarity matrix) and joining the corresponding sites or clusters to each other as a base for construction of dendrogram.
4- Updating the similarity / dissimilarity matrix according to similarities/dissimilarities between emerging cluster in Step (3) and other sites or clusters.
5- Repeating Steps (3) to (4) for (n-1) times till all individual sites and clusters are converted into final cluster.

The only difference among the various agglomerative techniques is in the Step (4). Methods of Single linkage and Complete Linkage are the extremes among the hierarchical methods as the Single linkage merges the clusters through nearest members while Complete Linkage merges the clusters through farthest members. Euclidean distance is currently used as the dissimilarity measure that gives optimum dissimilarity in
spherical vector space. It is defined as:

$$D_{j,k} = \sqrt{\sum_{i=1}^{p} (C_{j}^{i} - C_{k}^{i})^{2}}$$  \hspace{1cm} (1)

where, $D_{j,k}$ is the Euclidean distance between sites $j$ and $k$. $C_{j}^{i}$ and $C_{k}^{i}$ are the $i^{th}$ variable at the sites $j$ and $k$. Generally $C(.)$ can be shown as a vector of variables $(x_{1}, x_{2}, ..., x_{p})$.

Equations (2) and (3) give the distance between clusters $(p,q)$ and $(p',q')$ in Single Linkage and Complete Linkage respectively and can be written as:

$$D_{(p,q),(p',q')} = \min[D_{(p,q),p'}, D_{(p,q),q'}] = \min[D_{p,p'}, D_{q,p'}, D_{p,q'} D_{q,q'}]$$  \hspace{1cm} (2)

$$D_{(p,q),(p',q')} = \max[D_{(p,q),p'}, D_{(p,q),q'}] = \max[D_{p,p'}, D_{q,p'}, D_{p,q'} D_{q,q'}]$$  \hspace{1cm} (3)

An appropriate method between the two can be chosen according to the resulting dendrograms, which show the quality of the separation of clusters and the finally derived clear clusters can be used for delineation of homogeneous regions. It can be seen from the Eq. (2) that the single linkage method intends to contract the cluster space, by choosing the minimum distance between clusters and this property can sometime lead to production of chain-linked clusters. So, elongated clusters are the likely feature of this method. Contrary to this, the complete linkage intends to dilate the cluster space by choosing maximum distance between clusters. So, in this method each site acts like a nucleus for a new cluster instead of joining one another like a chain. Since single linkage method has high compatibility with high-density clusters is therefore capable of separating non-elliptical clusters. Both methods are however, completely monotonic and there is no cross over on their dendrograms. So, recognition of homogeneous regions by both of these methods is easy. While both the methods have high capability to aggregate sites that can be contained in naturally spherical clusters, complete linkage method does not have a good capability to contain objects or sites that follow naturally eccentric or elongated cluster because of space dilation property. To sum up it can be said that in spite of similarities in procedure, these methods may yield perfectly different results.

**Test of Homogeneity and Selection of Best- Fit of Distribution**

$H_{1}$-Statistics method of homogeneity is one of the latest techniques available for establishing regional homogeneity using L-Moments (Hosking and Wallis [5]). The method is based on the evaluation of a standardised value ($H_{1}$) which is computed using say $V$ {real life say weighted L-Coefficients of variation (L-Cv)s} and the average and standard deviation of simulated L-Cvs ($Vs$) for the chosen region. A region is considered,
‘acceptably homogeneous’ if $H_1 < 1$; ‘possibly homogeneous’ if $1 < H_1 < 2$; and ‘definitely heterogeneous’ if $H_1 > 2$. Necessary equation for computation of homogeneity index ($H_1$) with say weighted L-Cv values can be written as:

$$H_1 = \frac{(V - \mu_V)}{\sigma_V} \quad (4)$$

where, $V$ = weighted L-Cv values and $\mu_V$ and $\sigma_V$ respectively are the mean and standard deviation of simulated $N_{sim}(=500)$ values of $V$.

In addition to the above, two additional measures $H_2$ and $H_3$, based on L-Cv/L-Ku and L-Cs/L-Ku ratios can also be computed separately in place of $V$, with the similar connotations of homogeneity as above.

Once an acceptably homogeneous or possibly homogeneous region is established, selection of an adequately fitted distribution is carried out which is based on the distance of the point $(z_{Dist})$ plotted using regional L-Cv and L-Kurtosis (L-Ku) values from the curves of several three parameter distributions on a L-Cv and L-Ku diagram. All distributions whose absolute $Z$ value is less than 1.64 qualified as a possible candidate, with the lowest value qualifying for the best fit distribution. Thus, $Z_{Dist}$, can be computed from:

$$Z_{\text{Dist}} = \frac{\tau^\text{Dist} - \tau^4}{\sigma^4}$$  \quad (5)

where, $\bar{\tau}^4$ = average L-Ku value computed from the data of the region, $\beta^4$ = bias in L-Ku values, $\tau^4$ computed from the original data, $\tau^\text{Dist}$ = average of L-Ku value computed from simulation of a fitted distribution, $\sigma^4$ = standard deviation of L-Ku values obtained from simulated data.

Once the regions have been delineated and the underlying distribution representing the region identified, the next most important step is to obtain flood quantiles at different recurrence intervals ($x_T$) at different gauging stations which are then standardized to $X$ by dividing them with their respective annual mean flood values ($\bar{x}$). Plot of $X = (x_T / \bar{x})$ vs $Y = [-\ln(-\ln(1-1/T))]$ gives the growth curve of the region.

Development of Equations for Estimation of Annual Average Flood ($\bar{x}$) for Ungauged Catchments in the Region.

Estimation of quantiles at ungauged locations within a homogeneous region using the growth curve requires an apriori estimate of the mean peak flood ($\bar{x}$) at those locations; such that standardised flood quantiles obtained using the regional growth curve for the desired return periods can be converted to real life values. For developing such equations for each of the identified regions the following general form of regression model have been considered
\( \bar{x} = a \cdot A^b \cdot L^c \cdot S^d \cdot R_d^e \)  

where A, L, S, Rd are the various catchment characteristic related to Area, Main Channel Length, Slope and Rainfall in Design Duration and a, b, c, d and e are the coefficients of regression.

**STUDY AREA AND DATA**

The area under consideration is situated in central and south central part of India, between latitudes 13° 7' - 24° 57' N and longitudes 69° 8' - 87° 2' E, extending over 104,1679 km² area which approximately covers a third of the total area of the country. It is covered with both mountainous as well as coastal belts thus experiencing a varying elevation from 0 to 1350 m. It also experiences mean annual rainfall (MAR) of about 1000 mm. Main rivers flowing across the study area are: The Baitarani, Brahmani Mahanadi, Mahi, Sabarmati, Saraswati, Narmada, Tapi, Godavari, Krishna, and the Penner. Totally, 88 number of bridge sites where peak flood observations made by Railway Design and Standards Organization (R.D.S.O) at Lucknow have been collected for 35 years together with other relevant basin characteristic data such as: mean annual peak flood data \((\bar{x})\) (in m³/s), catchment area \((A, \text{ km}^2)\), main stream length \((L, \text{ km})\) weighted slope of main stream \((S, \text{ m/km})\), design duration \((T_d, \text{ hour})\), rainfall in design duration \((R_d, \text{ mm})\), and average \(\Phi\) - index \((\text{mm/hour})\), which expresses the total loss of rainfall containing the geological properties of a catchment. So, the variable vector \(W\) used for recognition of pattern within the sites can be represented in the form:

\[
W = (S_{SPMF}, S_L, S_{RS}, S_{RRDTD}, S_{PHI})
\]

where, \(S_{SPMF} = \text{Standardised Specific Mean Flood} \left( \frac{(\bar{x})}{A}/\sigma((\bar{x})/A) \right)\) and similarly \(S_L = \text{Standardised Length}; S_{RS} = \text{Standardised Root Main Stream Slope}; S_{RRDTD} = \text{Standardised Rainfall in Design Duration to Design Duration and } S_{PHI} = \text{Standardised Phi-Index.}

**RESULTS OF ANALYSES**

The resulting dendrograms developed using the single and complete linkage procedure (Eq. 2 & 3) have been illustrated in Figs. (1) and (2) respectively. Comparing the two dendrograms it is evident that complete linkage method has yielded less number and much clear-cut clusters than single linkage. Consequently, the result of complete linkage method was selected for delineation of homogeneous regions.
To start with, instead of just looking at 3 major clusters, which within themselves could be mosaic, many sub-regions were identified which were then joined with adjacent sub-regions or individual sites to form continuous large scale regions in such way that the finally emerging regions were either homogeneous or were not too heterogeneous and have been shown in Fig. (3) as regions A, B, C, D, E, F, G, and H. For each of these regions, \( H_1 \)-statistics computed using Eq. (4) have been tabulated in Table 1.

Table 1. Homogeneity test values of regions

<table>
<thead>
<tr>
<th>Region</th>
<th>Sites Identification Number</th>
<th>No. of sites</th>
<th>( H_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1 16 18 31 32 33 34 35 36 37 38 39 40 41 42 43 66 70</td>
<td>18</td>
<td>1.49</td>
</tr>
<tr>
<td>B</td>
<td>2 3 4 5 8 10 11 12 68 69 73 75</td>
<td>12</td>
<td>2.80</td>
</tr>
<tr>
<td>C</td>
<td>6 7 9 13 14 15 17 67 71 72 74 76</td>
<td>12</td>
<td>2.53</td>
</tr>
<tr>
<td>D</td>
<td>77 79 80 81 82 83 84 85 87 88 78 86</td>
<td>12</td>
<td>0.73</td>
</tr>
<tr>
<td>E</td>
<td>53 57 59 62 63 65</td>
<td>6</td>
<td>1.85</td>
</tr>
<tr>
<td>F</td>
<td>55 56 60 61 64</td>
<td>5</td>
<td>2.99</td>
</tr>
<tr>
<td>G</td>
<td>44 45 46 47 48 49 50 51 52 54 58</td>
<td>11</td>
<td>1.73</td>
</tr>
<tr>
<td>H</td>
<td>19 20 21 22 23 24 25 26 27 28 29 30</td>
<td>12</td>
<td>-.69</td>
</tr>
</tbody>
</table>

From the homogeneity values, it can be seen that, regions A, D, E, G, and H are acceptably homogeneous while regions B, C, and F show somewhat heterogeneity. Since small departure from homogeneity range does not negate the benefit of regionalization (Cunnane, [2]), these regions were finally accepted. Values of goodness-of-fit measure \( Z_{\text{Dist}} \) for each region have been computed to obtain the best-fit distribution using Eq. 5 and values tabulated in Table 2.
Figure 3. Resulting homogeneous regions from the application of Complete Linkage Method.

Table 2. Regional parameters of the best-fit distributions and equations for estimation of mean annual floods for ungauged catchments within the identified regions.

<table>
<thead>
<tr>
<th>Region</th>
<th>Z^Dist. value</th>
<th>Dist.</th>
<th>Regional Parameters</th>
<th>Regional Relation</th>
<th>Std Error of Coeffs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.68</td>
<td>GN</td>
<td>α = 0.82, β = 0.61, γ = -0.54</td>
<td>Χ = 1.568 A^0.936</td>
<td>0.13</td>
</tr>
<tr>
<td>B</td>
<td>0.17</td>
<td>GN</td>
<td>α = 0.88, β = 0.55, γ = -0.42</td>
<td>Χ = 1.834 A^0.80</td>
<td>0.12</td>
</tr>
<tr>
<td>C</td>
<td>0.50</td>
<td>GN</td>
<td>α = 0.86, β = 0.57, γ = -0.45</td>
<td>Χ = 10.15 A^0.652</td>
<td>0.09</td>
</tr>
<tr>
<td>D</td>
<td>0.79</td>
<td>PTIII</td>
<td>α = 1.00, β = 0.78, γ = 1.48</td>
<td>Χ = 8.69 A^0.597</td>
<td>0.07</td>
</tr>
<tr>
<td>E</td>
<td>0.24</td>
<td>PTIII</td>
<td>α = 1.00, β = 0.69, γ = 1.13</td>
<td>Χ = 594.26 S^-0.388</td>
<td>0.15</td>
</tr>
<tr>
<td>F</td>
<td>-0.05</td>
<td>GN</td>
<td>α = 0.79, β = 0.63, γ = -0.6</td>
<td>Χ = 2.307 A^0.597</td>
<td>0.24</td>
</tr>
<tr>
<td>G</td>
<td>0.28</td>
<td>GN</td>
<td>α = 0.76, β = 0.63, γ = -0.67</td>
<td>Χ = 9.39 A^0.53 L^0.46</td>
<td>0.07, 0.15</td>
</tr>
<tr>
<td>H</td>
<td>0.44</td>
<td>GN</td>
<td>α = 0.84, β = 0.65, γ = -0.47</td>
<td>Χ = 4.41 A^0.646</td>
<td>0.19</td>
</tr>
</tbody>
</table>

It is evident from this table, that the majority of regions namely A, B, C, F, G, and H follow the Generalized Normal distribution (GN) while the regions D and E follow the Pearson type III (PTIII) distribution. For each of these regions, parameters for the best-fit distributions have been computed by the method of L-Moments and tabulated in Table 2. These parameters however have been used to develop the growth curve for each region. Fig 3 illustrates the growth curves of regions. Regional equation relating (Χ) and relevant catchment characteristics for each region has been developed as per Eq. (5) and the relevant coefficients and standard error of these coefficients have also been given Table 2.
CONCLUSIONS

In clustering the flood sites or objects, it is difficult to state which method is the best or better, because the nature of clusters and their shape are unknown initially. So, it is often a practice to use several methods and compare their results for arriving at optimum results or making a suitable choice. According to dendrograms the connection distances of clusters in single linkage method are smaller than that of complete linkage, which means that separation of clusters in complete linkage is better than that of single linkage. This returns to the space contraction property of the SLM contrary to the larger connection distances of clusters in CLM, which is based on the space dilation property. As a result of this, small changes in variables is likely to result in big changes in the results of single linkage method and this sort of sensitivity is not be observed in complete linkage method. However, based on the well separation of clusters and high homogeneity within the delineated hydrological regions, the complete linkage method seems to be a better option for delineating flood regions in Central India.

REFERENCES
