JFE JOURNAL OF FLOOD ENGINEERING
1(1) January–June 2010; pp. 51–66

FLOOD REGIONALIZATION USING A MODIFIED REGION OF INFLUENCE APPROACH

Saeid Eslamian

Dept. of Water, Isfahan University of Technology, Isfahan 84155-83111, Iran

ABSTRACT: During the last two decades, considerable efforts have gone into development of hydrologic regionalization procedures. The present investigation modifies the "Region of Influence"(RI) approach, based on statistical characteristics of the stream gauging stations in south eastern New South Wales (NSW), Australia. Statistical properties of flood data are included as the attributes in similarity distance algorithm instead of geographical location. The RI approach is modified by the determination of the criteria for the choice of a set of six statistical attributes of catchments. The choice of the threshold value of the similarity distances and the effect of attribute interaction are also discussed. The modified method is applied to determine RI of some streamflow gauging stations in NSW using three independent statistical attributes.

Keywords: Flood Regionalization, "Region of Influence" Approach, Statistical Hydrology, Classification, New South Wales.

1. BACKGROUND

1.1. Introduction

A common problem for hydrologists is the estimation of peak flow frequencies at locations with little or no flood records. The Index Flood method; multiple regression techniques; the canonical correlation analysis (Cavadias *et al.*, 2001), the "Region of Influence (RI) approach (Burn, 1990a, b; Zrinji and Burn, 1994), cluster analysis (Rao and Srinivas, 2006) and L-moments (Hosking and Wallis, 1993; Yue and Wang, 2004) have been applied to contiguous catchments and they delineated fixed regions. The resulting regions are sometimes not hydrologically homogeneous, especially if the spatial variability of the physiographic or hydrologic characteristics is large. The main assumption for the traditional methods is that flood data from different stations in a region should be independent. If intersite dependence between nearby gauging stations is present, traditional methods give unreliable results. A regionalization method that does not utilize the distinct boundaries between defined regions may overcome these problems.

The fundamental concept of the RI approach, first described by Acreman and Wiltshire (1987), is that there is no need for distinct boundaries between different regions and each reference site can be allowed to have unique set of sites which constitutes the "Region" for this

reference site. Burn (1990a) conducted an interesting study in an attempt to improve extreme flow estimates for sites in southern Manitoba, Canada. He described a methodology for transferring information from the surrounding stations to the reference site for improvement of single-site extreme flow estimates.

1.2. Associated Problems with Burn's Paper

Burn (1990b) employed two flood statistics consisting of the coefficient of variation (C_v) and specific mean annual flood ($\bar{Q}/A$) with combination of two physical attributes of catchments, latitude and longitude from the reference station, using annual flood series. It is perceived that the mean flood ($\bar{Q}$) is considered in both the flood statistics and physical attributes were included in the similarity distance algorithm; the physical attributes dominate over the interacted flood statistics. This problem decreases the dispersal of surrounding stations in relation to the reference site in the RI. Therefore, some dissimilar stations will be added to the set of similar sites in the RI. In this way, although the included sites in the RI nicely surround the reference site and although the RI's appropriately overlap, it may lead to some sites to have largely different C_v and $\bar{Q}/A$ included in their RI. Burn (1990, a, b) did not use C_s (coefficient of skewness) due to the associated sampling error. It is expected that the inclusion of an attribute with the same properties of C_s in the similarity distance algorithm would improve the RI approach. The sensitivity of the standardization of attributes and the threshold value of the similarity distance have not been also investigated.

In the present study, the RI approach is chosen to be developed. This approach is a new technique that transfers the relevant information from the dispersed catchments in the RI to their reference site. It is not necessary to have distinct boundaries between defined regions since each site has a potentially different set of stations included for the single-site estimation or improvement of extreme flow values.

2. STUDY AREA

For regionalization studies, stream gauging stations having low and high temporal variation of floods are preferable. The chosen study area is located in south-eastern New South Wales (NSW), Australia such that it includes the Great Dividing Range, which runs parallel to the east coast and drains both inland and seaward. The head waters of the most variable streams are located in the slopes of the Great Dividing Range (McMahon, 1982). It would be useful to be able to show the effect of catchment area (or in another form Q_{50}/A) on the derived homogenous regions. In general, it would be desirable to have a network with a full range of basin areas, all climate types, vegetation, landform and major relief. Altitude is largely variable in the east-west direction (0 to about 1500 m) for the region under investigation. Figure 1, is a modified picture of PC GLOBE (1990). The study area covers approximately 250,000 square kilometers having mainly a temperate moist climate (Figure 1). In the selected region for the present investigation, 58 stream gauging stations with more than 30 years of data have been candidate.

Figure 1: Situation of Study Area on the Australian Climate Map

3. METHODOLOGY

3.1. Development of a *RI* Approach

The RI approach is based on delineation of a "Region of influence" for each gauging station including the set of sites that are in proximity to the candidate station, where proximity is defined in terms of the selected attributes rather than geographical location. Proximity is calculated by similarity distance algorithm in a multi dimensional attribute space where the attributes are appropriate measures for the identification of stations with a similar extreme flow response. In this section, the objective is the development of RI approach for the improvement of flood quantiles. It is suggested that each candidate attribute should be studied separately

Table 1
Model Development Accomplished by the Author in Comparison with those for Burn (1990a)

No.	*Author*	*Burn*
1	Six candidates from statistics of partial flood series are evaluated to find most suitable flood stations for inclusion in the similarity distance algorithm.	Burn arbitrarily selected four attributes, $C_{v,}$ $\bar{Q}/A$, longitude and latitude, using annual series.
2	Two possible options for defining the threshold value of the similarity distance are determined.	None
3	Sensitivity of the threshold is evaluated.	None
4	A requirement of a scaling factor (standardization) is examined.	Burn used a scaling factor without examining it.
5	A weighting function is modified and includes the threshold value (T_r) and a power of 2 for the parameters.	Burn used a power 4 and THL instead of $T_{r,}$ as threshold value. THL $\geq T_r$.
6	A weighting coefficient for each of the attributes is determined.	None

before being combined with other attributes for inclusion in a similarity distance algorithm. This leads to the establishment of a set of attributes with different applied weight to form the basis for the distance measure. Then, two different options are presented for defining a threshold value of the similarity distances. A discussion is made on the requirement of a standardization technique for scaling the attributes. It should be noted that all candidate attributes are calculated on the basis of partial series. The model development and accomplishments which will be presented in this study are compared with Burn's work (1990a) in Table 1.

3.2. Attribute Selection

To ensure that values for the attributes can be estimated for each of the stations in the data set, it is necessary that attributes be limited to characteristics that are readily obtainable for a network of stations of different catchment size and diversity of other characteristics. As described before, Burn (1990a) employed two flood statistics consisting of the coefficient of variation (C_v) and specific mean annual flood ($\bar{Q}/A$) with a combination of two physical attributes of catchments, latitude and longitude from the reference stations, using annual flood series.

One of the problems associated with Burn's method was the interaction in selected catchment attributes. Furthermore, Burn (1990a, b) did not use the skewness as an attribute.

It is expected that the inclusion of this attribute or an attribute with the same properties in the similarity distance algorithm would improve the RI approach. There are two general types of attributes that could be employed in the similarity distance algorithm. A combination of statistical features and physical attributes in formation of the similarity distance algorithm should be important since there are many factors which influence the magnitude of the *T*-year flood at a particular site. The major advantage of the statistical measures of the peak discharge data at each stream gauge is that they can be provided more easily than those of the physiographic measures. Furthermore, gauging stations which are similar in statistical attributes could be anticipated to have similar extreme flow responses. In this study, the attention is focused on statistical attributes and physical attributes are leaved for further studies.

3.2.1. Probability Density Function

Many candidate attributes were examined with respect to their influence on the probability density function (PDF) of flood and consequently their influence on the extreme flow values. Figure 2 (a) and (b) demonstrate the PDF curves for two stations, 212011 and 212320, where the number of occurrences is expressed as a proportion of the total. The probability that any randomly selected flow will be less than a given value is given by cumulative distribution function (CDF) of that population, $F(q)$, (Figure 3).

Development of a flood frequency analysis procedure involves choosing a distribution which is considered to describe most adequately the available flood series. Log Pearson Type 3 was adapted as the most appropriate distribution for the partial series to the majority of stations in the region under investigation (Eslamian, 1995). It was fitted to 58 stations including stations 212011 and 212320 for which there is remarkable difference between their PDFs. Indirect

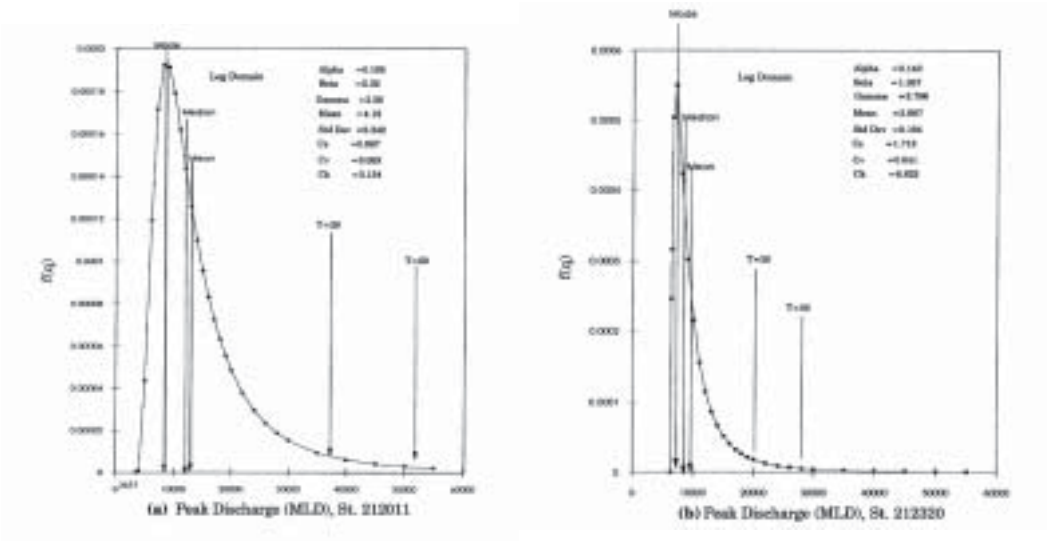

Figure 2: Probability Density Function

moments are used to estimate parameters of the Log Pearson Type 3 distribution. The efficiency of such moment estimators in small sample size (n = 31 and n = 23 for stations 212011 and 212320, accordingly) is discussed in Arora and Singh (1989). In comparing the two PDF curves to investigate candidate statistical attributes, it is clear that for stations 212011, sample values of C_v, Cs and standard deviation are respectively nearly 1.5, 0.5 and 1.5 time value for station 212320 (Figure 2). These differences for attributes between the stations cause remarkable variability in extreme flow values. As an example, at T = 20 and T = 50, discharges for station

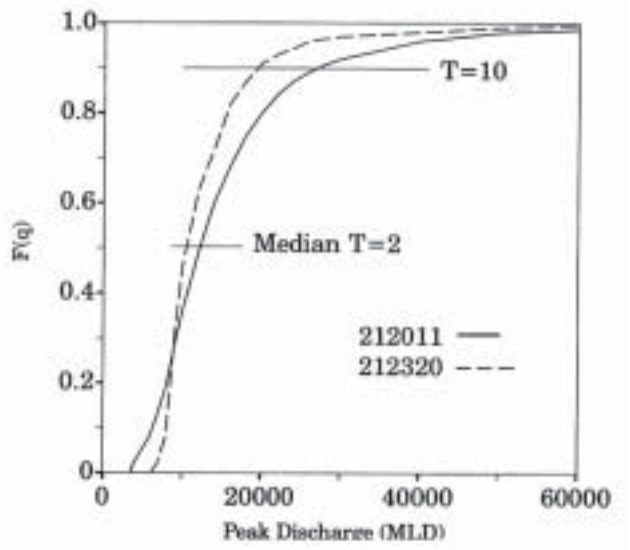

Figure 3: Cumulative Distribution Function for Partial Series

212320 are respectively 23900 Mega Liters per Day (MLD) and 33100 MLD, which are 65 per cent of those for station 212011. The four-fold differences in catchment areas (404 km^2 vs. 89.6 km^2, for stations 212011 and 212320) might also be a factor in difference in flood quantiles, particularly through the slope of the flood frequency curves. In flood studies, the estimation of extreme flow values for return period greater than 10 years is usually required. As shown in Figure 3, the frequency curve slope gently increases above the T = 10 boundary. This means that with a slight increase in $F(q)$, extreme flow values will increase much more. Therefore, there is a common difficulty in the estimation of the extreme flow for the long return intervals that are required in the design of many hydraulic structures. The starting point for the attributes' selection is to consider many statistical attributes which are basic in formation of flood growth curve. These are described in the following sections.

3.2.2. (α, β), (m–γ), γ and (μ–m) / (μ–γ)

The parameters of the Log Pearson Type 3 (LP3) distribution, α, β and γ can be described in terms of statistical characteristics of a sample as follows:

$$\beta = \left(\frac{2}{C_s} \right)^2 \tag{1}$$

$$\alpha = \frac{\sigma}{\sqrt{\beta}} = \frac{C_s \sigma}{2} \tag{2}$$

$$\gamma = \mu - \sigma\sqrt{\beta} = \mu\left(1 - 2\frac{C_v}{C_s}\right) \tag{3}$$

where μ is the mean, σ is the standard deviation, C_v is coefficient of variation and C_s is coefficient of skewness of the series. If we substitute the parameters in the probability density function of LP3 and solve them for m and γ, we obtain:

$$m - \gamma = \frac{4\sigma - \sigma C_s^2}{C_s(\sigma C_s + 2)} \tag{4}$$

In this equation, if $C_s = 0$, then $m - \gamma = \infty$ (Normal distribution). If we plot $m - \gamma$ vs. against C_s (Figure 4), it will be clear that not only skew but also standard deviation play an important role in flood frequency analysis.

Another dimensionless attribute, which is an attribute for inclusion in the similarity distance algorithm, can be written as:

$$\frac{\mu - m}{\mu - \gamma} = \frac{\mu - m}{\alpha\beta} \tag{6}$$

3.2.3. Coefficient of Variation

The variability of flood series can be characterized in dimensionless form that is defined as:

$$C_v = \frac{\sigma}{\mu} \tag{7}$$

which can transfer the standard deviation characteristics in a more appropriate form into similarity distance algorithm. Some hydrologist (Mosley, 1981, Wiltshire, 1986, Bhasker and O'Connor, 1989, Burn, 1990 a) defined homogenous region by the flood-related variables such as C_v.

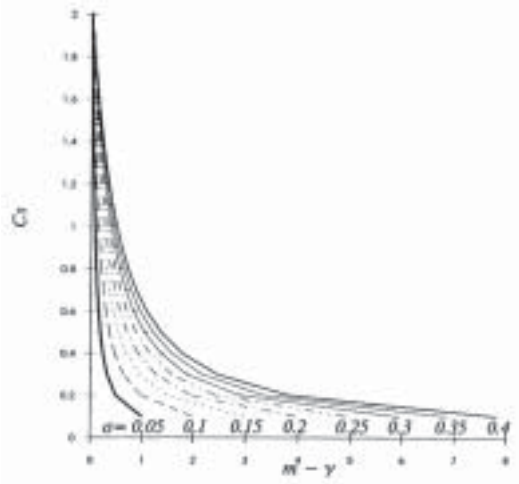

Figure 4: Theoretical Relationship between Some Attributes for LP3 Distribution

3.2.4. Coefficient of Skewness

The coefficient of skewness plays a very important rule in forming the PDF and flood frequency curves and can be used for measuring similarity between sites. A screening process is an important

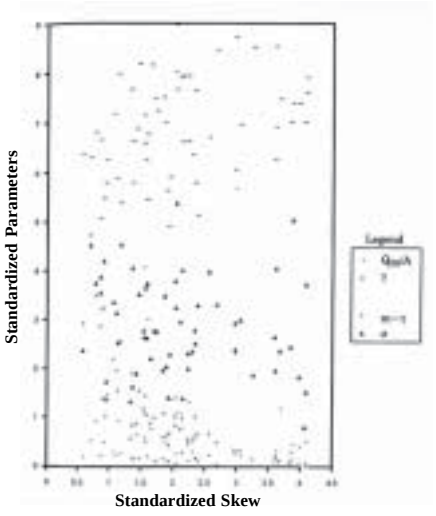

Figure 5: (a) The Screening Process of Some Candidate Attributes

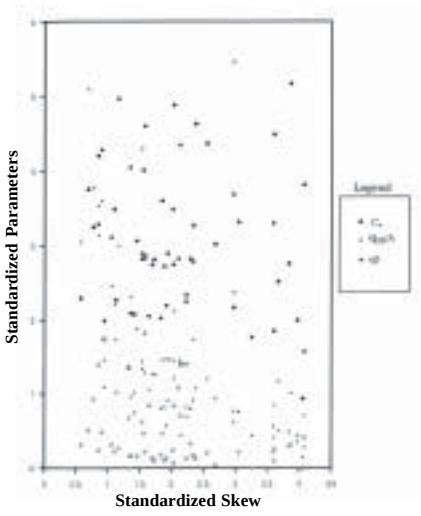

Figure 5: (b) The Screening Process of Some Candidate Attributes

tool for identification of the relationship between attributes. The bivariate plot of skew versus seven variables is shown in Figures 6 (a) and (b). Because the variables are of different magnitudes and are expressed in different unites, three is a requirement to standardize the value of variables before performing the screening process. A standardization technique used in this study involves dividing each candidate attribute by the standard deviation of this attribute for all the stations. As shown in Figures 5 (a) and (b), all data points are scattered and specific relations do not seem to exist between skew and the majority of the candidate attributes.

3.2.5. Specific 20 or 50-year Flood

Log Q_{20}/Log A and log Q_{50}/Log A describe the location of individual flood estimates. As Q_{20} and Q_{50} are highly correlated, it seems that one of them is sufficient to transfer the importance of extreme flow values. In this case, a 20-year flood is preferred to 50-year as it was used by some hydrologists (for example, Tasker, 1987). Burn (1990a) has used $\bar{Q}/A$ which is correlated with C_v. Because this attribute was again employed for calculating C_v, as shown in eq. 7, it is not recommended that $\bar{Q}/A$ be used in the present study. Some investigators (e.g. Riggs, 1990;) showed that C_v varies inversely with drainage area in regions they studied. For catchments investigated in the present study, there is negative, but not strong correlation between C_v and A ($R^2 = 0.38$).

3.2.6. Log Q_{10} / Log Q_2

Log Q_2 and *Log* Q_{10} are the logarithm of at-site flood values for 2-year and 10-year return periods. It is better to use this attribute in log domain as it reflects which leads it to have the same properties. *Log* Q_2 and *Log* Q_{10} can be written:

$$Log\ Q_2 = \mu + f_2\sigma \tag{8}$$

$$Log\ Q_{10} = \mu + f_{10}\sigma \tag{9}$$

Subtracting eq. (8) from eq. (9) gives:

$$Log\frac{Q_{10}}{Q_2} = \sigma(f_{10} - f_2) \tag{10}$$

In the other form, eq. (10) can be written:

$$\frac{LogQ_{10}}{LogQ_2} = \frac{1 + f_{10}C_v}{1 + f_2C_v} \tag{11}$$

It seems to be logical that left side of eq. (11) would be affected by σ less than eq. (10). Therefore, it is better to use $LogQ_{10}/LogQ_2$ than $Log(Q_{10}/Q_2)$ and Q_{10}/Q_2 for combination with other attributes that are correlated with σ (such as C_v).

3.2.7. Pearson Skewness

Pearson skewness (PSK) is a dimensionless coefficient that is inversely proportional to a standard deviation and is defined through:

$$PSK = \frac{\mu - m}{\sigma} \tag{8}$$

which can be transferred to:

$$PSK = \frac{\mu}{\sigma} - \frac{m}{\sigma} = \frac{1}{C_v} - \frac{m}{\sigma} \tag{9}$$

where m is median of the flood series.

As the previous attributes, (α, β), (m–γ) and γ are not dimensionless, we exclude them from attributes used in a similarity distance algorithm.

3.3. Combination of Attributes

An extensive range of various proximity measurements has been suggested by the Statistical Package for the Social Science, SPSS (2002) for classification investigations such as the grouping of stations into regions. The emphasis in this study is on describing proximity measures for a pair of sites by a set of attributes. The Minkowski metric (Gordon, 1981) is the appropriate algorithm which allows some weighting of attributes as follows:

$$\Delta_{jk} = \left[\sum_{i=1}^{p} W_i \left| O_{ji} - Q_{ki} \right|^{\lambda} \right]^{1/\lambda} \quad (\lambda > 0) \tag{10}$$

Δ_{jk} = metric distance from site j to site k.

p = the number of attributes that has been selected.

W_i = weighting coefficient, which implies proportional importance of attribute i with other attributes.

O_{ji} = standard value of attribute i for site j.

λ = a power such that high values cause more emphasis on the difference of a pair of sites for the value of a specific attribute.

The specific case λ = 1and λ = 2 were generally used in eq. 10:

$$\Delta_{jk} = \sum_{i=1}^{p} W_i \left| O_{ji} - O_{ki} \right| \tag{11}$$

$$\Delta_{jk} = \left[\sum_{i=1}^{p} W_i \left| O_{ji} - Q_{ki} \right|^{2} \right]^{1/2} \tag{12}$$

where eq. 11 and eq. 12 are City Block metric and Euclidean distance, respectively (Gordon, 1981). For the combination of attributes in the RI approach, the Euclidean distance (λ = 2) is used. For a set of six attributes, Eq. 12 is written as:

$$\Delta_{jk} = \left[\sum_{i=1}^{6} W_i \left| O_{ji} - O_{ki} \right|^2 \right]^{1/2} \tag{13}$$

The weighting coefficient (W_i) can be determined using the determination coefficient (Ri^2) between each attribute (i) and 100-year extreme flood values (real domain), as $W_i = R_i^2/R_1^2$, where $R_1^2 < \ldots < Ri^2 < \ldots < R_p^2$.

3.4. Standardization

The attributes usually have different units or magnitudes which are not comparable. The similarity measure is highly dependent upon the scales of measurement used. The common method for standardization is to equalize the standard deviation of attributes. Thus, each attribute must be redefined by dividing by the standard deviation of that attribute for all stations. This technique is used for the present study. Therefore, if the original series ($x_1, x_2, \ldots, x_n$) is to be standardized, the standardized series will be ($y_1, y_2, \ldots y_n$), where $y_i = x_i / \sigma_x$. Eslamian (1995) showed that the mean of standardized series is the inverse of the coefficient of variation of original series. He also showed that C_v has a great effect on obtained Δ_{jk}. This coincides with the ranges of C_v and $(\mu - m)/(\mu - \gamma)$ values. The difference of C_v values for stations 215004 and 4100705, for example, covers 53% of the whole of the range, which is approximately twice these for $(\mu - m)/(\mu - \gamma)$ (25%). Therefore, the standardization technique could transfer the importance of C_v to the Euclidean distance.

3.5. Threshold Definition

An important step in RI approach to regionalization is the judgment for defining a threshold value for the similarity distance. The threshold value defines a cutoff to exclude gauging sites from the region of the reference site. Any surrounding sites having Δ_{jk} lower than the threshold value are considered for inclusion in RI of the reference site. The options mentioned below suggest two possible definitions of the threshold:

Option 1

This option is to look for a breakpoint in the array of distance values, separately in each row of the sorted matrix (for each reference stations). Breakpoints indicate a possible threshold, but the main problem is to distinguish an appropriate breakpoint to be considered as a threshold.

Option 2

In this potion, a unique threshold for all stations will be defined. The average of each column of the sorted matrix makes up a single array [a matrix ($1 \times N$)], as shown in Figure 6.

There is a breakpoint at point *O*. This is clear, because the number of stations having Δ_{jk} = 2–3 is 1.5 times the number of stations having Δ_{jk} = 1–2, giving the change of slope. Thus, a value of threshold, T_r, equal to 2 might be best to be used.

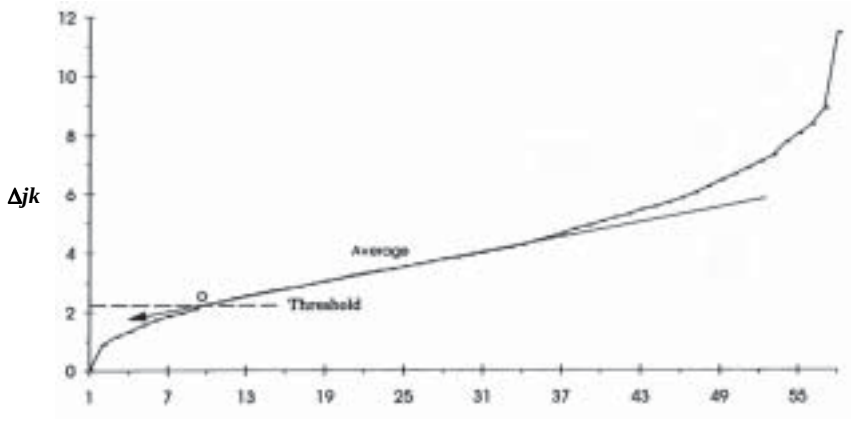

Figure 6: Average of Sorted Array of Δjk

3.6. Weighting Function

The objective of introducing a weighting function is to transfer the information of surrounding stations to the reference station in a delineated RI. The weighting function was modified as follows:

$$\psi_{jk} = \frac{T_r^2 - \Delta_{jk}^2}{T_r^2} \tag{14}$$

ψ_{jk} = weighting for site k with respect to a reference site j, if Δ_{jk}^2 is equal to or less than T_r.

T_r = the threshold value.

The power of 2 determines the rate of increase of ψ_{jk} for sites, in that similarity distance from the reference sites is decreased.

4. APPLICATION OF MEHODOLOGY

4.1. Attribute Selection

The foundation of RI approach is the selection of appropriate attributes which should be inserted into the similarity distance algorithm for defining the RI of each site. The inclusion of a set of diverse and independent attributes for similarity distance can increase transfer of information from surrounding stations to the reference site. A correlation matrix for the six attributes chosen in this chapter is shown in Table 2. It can be seen that some attributes interact. There is a high correlation between the coefficient of skewness (C_s) and $(\mu - m) / (\mu - \gamma)$, since the coefficient of skewness is proportional to $(\mu - m)$. Thus, as shown in Table 3, there is not a strong correlation between three attributes chosen in sections 3.2.2 to 3.2.7. Figure 7, demonstrates a 2-dimensional plot for these attributes for the catchments used in this study.

Table 2
Correlation Coefficient between Chosen Attributes

	C_v	C_s *Log A*	*Log Q50/* (μ–γ)	*(μ–m)/*	*PSK*	*Log Q10/ Log Q2*
C_v	1.0					
C_s	-.09	1.0				
*LogQ*50/*LogA*	0.27	-0.05	1.0			
	-0.06	0.98	-0.09	1.0		
PSK	-0.88	0.2	-0.36	0.21	1.0	
*LogQ*10/*LogQ*2	0.88	0.13	0.24	0.13	-0.75	1.0

Table 3
Correlation Coefficients between Chosen Attributes

	C_v	*LogQ50/LogA*	*(μ–m)/(μ–γ)*
C_v	1.0		
*LogQ*50/*LogA*	0.15	1.0	
(μ–m) / (μ–γ)	0.02	-0.15	1.0

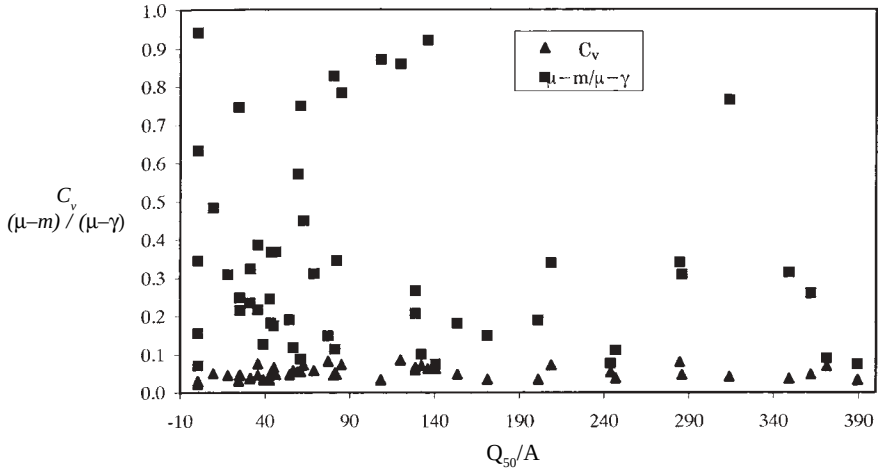

Figure 7: Screening Process of the Attributes

4.2. Comparison with the 6-Attribute RI

In this section, a set of six attributes is compared with a set of three relatively independent attributes to investigate their contribution to the identification of RI for sites. Station 219003 (Morans Crossing on Bemboka River) and 410534 (Upstream of Happy Jacks Pondage on Happy Jacks River) are two quite different cases with their RI's shown respectively, in Figures 9 and 10. These figures are plotted for both 3 and 6-attribute RI's.

The majority of stations that are included in the RI of site 219003 using the 6-attribute RI

approach are geographically different from the coastal site 219003 while in the case of 3-attribute RI, sites included in the RI display more similarity with the reference station 219003. Therefore, it can be said that the 6-attribute RI approach lead to define an unreal RI for some sites due to correlation between attributes. But, in the case of site 410534, although there are 12 sites in common with 3 and 3 and 6-attributes RI approaches, using the 6-attribute RI approach, two included coastal sites 215004 and 219003 are physically different from the other included stations.

The interesting point is that the C_v and $(\mu - m) / (\mu - \gamma)$ values for both stations 219003 and 410534 are close together, as shown in Table 3.

Table 3
Representation of the Three Attribute Values for Two Extreme Cases of the Stations

Gauging Stations	C_v	*Q50/A*	$(\mu - m) / (\mu - \gamma)$
219003	0.046	286.1	0.307
410534	0.046	81.96	0.344

Therefore, the *Q*50/*A* attribute is the cause of different situations for station 219003 and 410534 (Figures 9 and 10), because the attribute PSK, *Cs* an $LogQ_{10}$ / $LogQ_2$ have similar properties with C_v and $(\mu - m) / (\mu - \gamma)$. Value of 286.1 for *Q*50/*A* in the case of station 219003

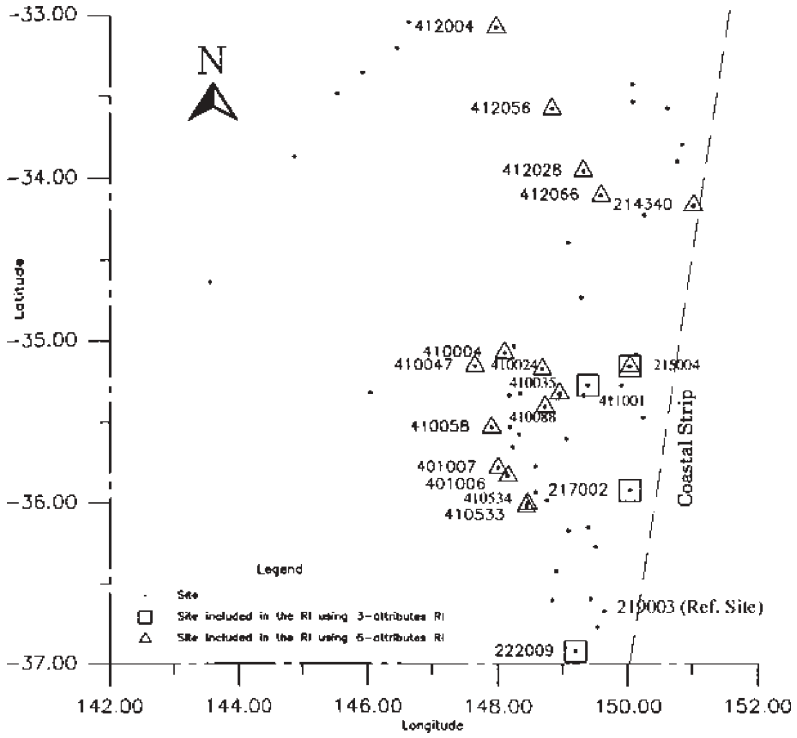

Figure 9: Comparison of the RI for Site 219003 between 3-attribute an 6-attribute RI

is a high value in comparison with other stations. This high value for $Q50/A$ has a dominant effect on the combination of three attributes, but in a combination of six attributes, this high value is covered by three extra attributes, because these three extra attributes have dissimilar properties to $Q50/A$. Therefore, for the reference station 219003, in a combination of six correlated attributes for the Euclidean distance, more stations with less similarity are included in the RI in comparison with the 3-atribute RI, as shown in Figure 9.

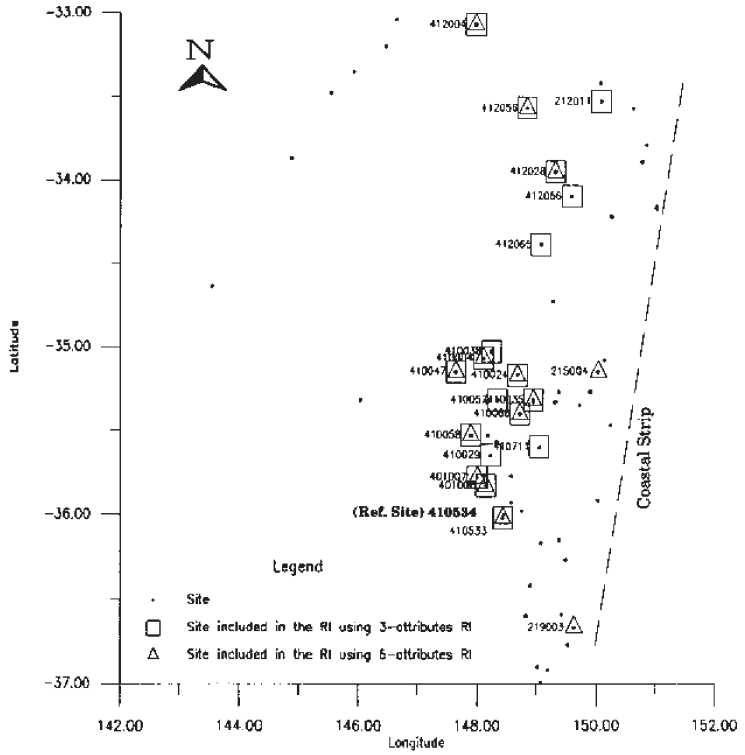

Figure 10: Comparison of the RI for Site 410534 between 3-attribute an 6-attribute RI

5. CONCLUSION

A value of 2 for the threshold value of the similarity distances on a set of 58 sites in south-eastern NSW can be compared with that of 1.8 that Burn (1990a) applied to a set of 91 streamflow gauging stations in southern Manitoba, Canada. This different can be explained in the type and number of attributes employed in both studies .Work on the Canadian stream used four attributes, the coefficient of variation (C_v), specific mean annual flood ($\bar{Q}/A$), longitude, and latitude, in comparison with six attributes, C_v, C_s, $LogQ_{50}/LogA$, $(\mu - m)/(\mu - \gamma)$, PSK and $LogQ_{10}/LogQ_2$, for the present study. This is due to the nature of eq. 12, involving a summation from $i = 1$ to p. In addition, including the weighting coefficient (W_i) in the Euclidean distance algorithm caused a weighting greater than 1 for most of the attributes, while W_i was not considered in Burn's distance metric (1990a).

Engineering judgment is necessary for an identification of the threshold value (T_r) of the similarity distance. Choice of large value for T_r decreases the homogeneity of a RI and increases the number of streamflow gauging stations included. It is better to say T_r is a compromise between the quantities of station's data surrounding the reference site and the quality of data in view of similarity in physiographic and statistical characteristics.

$LogQ_{50}/LogA$ is found as an important parameter not only for flood frequency analysis but also for inclusion in Euclidean distance algorithm in RI approach. A combination of extreme flow value and drainage area was used in previous works (Mosley, 1981; Wiltshire, 1986; Burn, 1990a).

Considerable amount of attention should be given to the specification of the appropriate weighting coefficient for attributes. It is considered that the weighting coefficients compared in section 3.3 for application to the similarity distance algorithm reflect the relative importance of attributes in relation to the one hundred year flood. However, as Moss (1968) noted, classification is found to be little affected by sight changes in the weighting coefficients.

It is finally concluded that there is a strong probability that the attributes interact. The effect of two attributes, which have the identical properties, equals to the effect of one those with a weighting coefficient of 2, when they are used in the similarity distance algorithm. Thus this suggests that fewer attributes with independent properties should be used. In sections 4.2, three independent attributes were employed in the similarity measure algorithm and, screening sites to be included in the RI was demonstrated due to the existence of a strong correlation between the attributes in the 6-attribute RI.

6. ACKNOWLEDGEMENTS

The author wishs to thank Emeritus Professor David H. Pilgrim and late David Doran for their invaluable comments on this manuscript.

References

[1] Acreman, M. C. and S. E. Wiltshire, Identification of Regions for Regional Flood Frequency Analysis (Abstract), EOS Trans. Amer. *Geophys.* Union, **68**(44) (1987) 1262.

[2] Arora K. and V. P. Singh, A Comparative Evaluation of the Estimators of the Log Pearson Type 3 Distribution, *Journal of Hydrology,* **105** (1989) 19-37.

[3] Bhasker, N. R. and C. A. O'Connor, Comparison of Method of Residuals and Cluster Analysis for Flood Regionalization, *Journal of Water Resources Planning & Management*, **115**(6) (1989) 793-808.

[4] Burn, D. H., An Appraisal of the "Region of Influence" Approach to Flood Frequency Analysis, *Hydrological Science Journal,* **35**(2) (1990a) 149-165.

[5] Burn, D. H., Evaluation of Regional Flood Frequency Analysis with a Region of Influence Appraoch, *Water Resources Research*, **26**(10) (1990b) 2257-2265.

[6] Cavadias, G. S., Ouarda, T. B. J. M., Bobee, B. and C. Girard, A Canonical Correlation Approach to the Determination of Homogenous Regions for Regional Flood Estimation of Ungaged Basins, *Hydrological Sciences Journal,* **46**(4) (2001) 499-512.

[7] Eslamian, S. S., *Regional Flood Frequency Analysis using New Region of Influence Approach*, Ph.D. Thesis, University of New South Wales, Australia, (1995) 411.

[8] Gordon, A. D., *Classification, Monographs on Applied Probability and Statistics*, Chapman & Hall, New York. (1981).

[9] Hosking, J. R. M. and J. R. Wallis, Some Statistics useful in Regional Frequency Analysis, *Water Resources Research,* **29** (1993) 271-281.

[10] McMahon, T. A., World Hydrology: Does Australia Fit? Hydrology & Water Resources Symposium, *Inst. Engrs Aust.,* Natl. Conf. Publ. No. **82**(2) (1982) 1-7.

[11] Mosley, M. P., Delimitation of New Zealand Hydrology Regions, *Journal of Hydrology,* **49** (1981) 173-192.

[12] Moss, W. W., Experiments with Various Techniques of Numerical Taxonomy, *Syst. Zoology,* **17** (1968) 31-47.

[13] PC GLOBE INC., *PC GLOBE Software,* Version 4, Tempe, AZ, USA. (1990).

[14] Rao, A. R. and V. V. Srinivas, Regionalization of Watershed by Hybrid-Cluster Analysis, *Journal of Hydrology,* **318** (2006) 37-56.

[15] Riggs, H. C, Estimating Flow Characteristics at Ungaged Sites, Regionalization in Hydrology, *Proc. of the Ljubljana Symposium,* IAHS Publ. No. 191 (1990) 159-169.

[16] SPSS Inc., *SPSS-x User's Guide,* McGraw-Hill, New York. (2002).

[17] Tasker, G. D., A Comparison of Methods for Estimating Low Flow Characteristics of Streams, *Water Resources Bulletin,* **23**(6) (1987) 1077-1083.

[18] Wiltshire, S. E., Identification of Homogenous Regions for Flood Frequency Analysis, *Journal of Hydrology,* **31**(3) (1986) 321-333.

[19] Yue, S. and C. Y., Wang, Possible Regional Probability Distribution Type of Canadian Annual Streamflow by L-moments, *Water Resources Management,* **18** (2004) 425-438.

[20] Zrinji, Z. and D. H. Burn, Flood Frequency Analysis for ungaged Sites using a Region of Influence Approach, *Journal of Hydrology,* **153** (1994) 1-21.