The canonical correlation approach to regional flood estimation

G. S. CAVADIAS
McGill University, Montreal, Canada

Abstract A method of estimating the maximum annual floods of a basin using canonical correlation analysis between the basin characteristics and the characteristics of the distribution of the floods is proposed in this paper. The method helps to determine the degree of similarity between the groupings of basins in the spaces of the respective canonical variables of the basin and the flood characteristics and provides a basis for the determination of the homogeneous regions in the space of basin characteristics which takes into account their relationships with the flood characteristics. The proposed method is used to estimate the quantiles of the distribution of the maximum annual flood in the province of Newfoundland in Canada.

L'approche de correlation canonique pour l'estimation régionale des caractéristiques des crues

Résumé Le présent article propose l'estimation de la crue annuelle maximale des bassins au moyen de la corrélation canonique avec leurs caractéristiques physiographiques. La méthode permet de déterminer le degré de similitude entre groupes de bassins dans les espaces des variables canoniques des caractéristiques physiographiques et celles des crues des bassins et partant de là, elle fournit une base pour la détermination de régions homogènes dans l'espace des caractéristiques physiographiques en tenant compte de leurs relations avec les caractéristiques des crues. La méthode est utilisée pour l'estimation des quantiles de la distribution de la crue maximale annuelle de la Province de Terre Neuve.

INTRODUCTION

The estimation of the hydrological variables of a basin in the case of non-existent or limited data can be improved upon by taking into account relevant information from other basins, provided that they constitute a "hydrologically homogeneous region".

Homogeneous regions were originally defined in the space of geographical coordinates. However this definition has the disadvantage that geographically contiguous basins may not by hydrologically similar (Linsley, 1982; United States Water Resources Council, 1988). As a result, a number
of researchers (Mosley, 1981; Gottshalk, 1985; Wiltshire, 1986) were led to define the homogeneous regions in the space of flood-related variables, e.g. specific mean annual flood and coefficient of variation using multivariate methods such as cluster and discriminant analysis. In order to estimate the flood quantiles of an ungauged basin it is necessary to assign it to a group of basins or “region” on basis of available physiographic and other basin variables and therefore this set of regions must be compared to that established on the basis of flood variables.

The reasoning on which the above approach is based is that the basic hydrological studies for a large area (continent, country, province, state, etc.) are carried out by government departments which publish reports with maps of the hydrologically homogeneous regions, multiple regression equations and dimensionless regional probability distributions to be used by the design engineer to estimate the flood probabilities of a gauged or ungauged basin (e.g. Panu et al., 1984). The same logic underlies the recommendation of a standard type of probability distribution for the analysis of maximum floods such as the log-Pearson III distribution in the United States (United States Water Resources Council, 1977) or the generalized extreme value distribution in the United Kingdom (Natural Environment Research Council, 1975). This approach is also similar to that used in structural engineering where committees of government specialists prepare building codes which include tables of allowable stresses and live loads for the use of designers of individual buildings. It must be considered, however, that the estimation of the maximum annual flood of given exceedance probability is a much more difficult problem with more serious consequences in case of failure than the problems of routine structural building design. Consequently, the estimation of this probability by simply identifying the "region" to which the basin belongs, substituting the values of its physiographic variables into a pre-estimated regression equation (or using a pre-specified regional dimensionless probability distribution) is a rather over simplified approach to a very complex problem which requires more detailed study.

HYDROLOGICALLY HOMOGENEOUS REGIONS

The subdivision of a large area (continent, country, province, state, etc.) into homogeneous regions for a given purpose (Grigg, 1976) is a standard tool in the geophysical sciences. Maps of climatic classifications (e.g. Griffiths & Driscoll, 1982), mean annual runoff (Lvovich, 1973) and regional water yield (WMO, 1986; Falkenmark, 1976) give basic information about the hydrological conditions of large areas. It must be noted, however, that the mapping of hydrological variables in the form of isolines or homogeneous regions becomes less feasible for smaller areas and shorter time periods over which the variables are averaged. For example, maps of hydrological variables such as maximum unit spring runoff (WMO, 1977), statistics of flood distributions (WMO, 1977) or parameters of flood distributions such as the generalized skew coefficient (United States Water Resources Council, 1977) are based on considerable smoothing which reduces their value in the case of specific basins.

In the case of maximum annual floods, geographically defined
homogeneous regions are convenient for the design engineer. However, "geographical proximity is no guarantee that the distributions of normalized maximum floods are identical or almost so" (Cunnane, 1986). Another research (Wiltshire, 1986) found that "geographical regions do not display as much internal homogeneity of flood frequency behaviour nor as much between-region homogeneity as regions defined in the space of flood characteristics such as normalized annual flood and coefficient of variation". In addition, tributary basins which, by definition, belong to the same geographical region do not necessarily have the same flood-producing properties as the main basin.

The determination of "regions" in the space of flood-related variables is usually carried out by cluster analysis, a method that sets out to discover "natural clusters" (Dillon & Goldstein, 1984) under the assumption that such clusters exist. However, the existence of natural clusters in the space of flood-related variables cannot be taken for granted without testing (Rogers, 1974; Dubes & Zeng, 1987). If such clusters do not exist, the final set of "regions" will depend on the clustering method used and the initial assumptions made.

A recent example of this approach is found in the previously cited paper by Wiltshire (1986) in which the author seeks to delineate homogeneous regions for the estimation of floods in the United Kingdom. According to Wiltshire, "the required number of groups (regions) is arbitrarily chosen and the collection of basins is (initially) randomly divided into similar sized groups".

The iterative relocation algorithm used in the study (Gordon, 1981) automatically allocates basins to groups according to a specified nearness criterion. An examination of Fig. 1 and 2 taken from Wiltshire's paper shows that there are no "natural clusters" in Fig. 1 and therefore the final set of
clusters or "regions" in Fig. 2 depends on the method and the initial assumptions made concerning the number and location of the homogeneous regions. Thus, if the "regions" are defined in the space of variables related to the distribution of floods the design engineer must use a more or less arbitrary subdivision of the set of gauged basins into homogeneous regions. This raises a fundamental question concerning the concept of "region". As long as regions are defined on a map, their reality can be justified on the basis of general location, topography, weather regimes, geomorphology, geology and other intuitively understandable factors. This is, however, not the same for regions defined in the space of flood-related variables.

In view of this difficulty, proponents of this latter method seek to relate these flood-related regions to regions in the spaces of geographical coordinates and basin characteristics using a subjective approach (Mosley, 1981; Gottschalk, 1985; Wiltshire, 1986). From the above, we can conclude that once the design engineer leaves the intuitively clear approach of geographical regionalization, he faces a number of problems concerning the reality and use of regions in the spaces of flood and basin variables. It is apparent, then, that a reassessment of the definition and purpose of "hydrologically homogeneous regions" is necessary in order to avoid an arbitrary subdivision of large areas into sets of homogeneous regions which depend on the methodology used and to allow the objective linking of the classification of basins in the spaces of flood and basin variables.

The method of flood estimation by canonical correlation (Cavadias, 1989) is a first step in this direction. An outline of the steps of this method is shown in Fig. 3. The $p$ basin variables $X_1, ..., X_p$ and the $q$ quantiles ($Q_2, Q_9, ..., Q_{100}$) or ($Q_2/\bar{Q}, ..., Q_{100}/\bar{Q}$) of the fitted distribution are used in a canonical correlation analysis which produces the two sets of canonical variables $(V_1, ..., V_p)$ and $(W_1, ..., W_q)$. In this method, the examination and testing of similarities and groupings of basins is carried out in the spaces $(V_1, ..., V_p)$ and $(W_1, ..., W_q)$ respectively. If the basin variables are good

![Fig. 2](image_url)
The canonical correlation approach to regional flood estimation

Predictors of the flood quantiles, the patterns of points in these two spaces will be similar. Given the high correlation between computed quantiles, the first two pairs of canonical variables \((V_1, V_2)\) and \((W_1, W_2)\) are likely to be sufficient for representing the relations between the basin and flood variables. Figures 4 and 5 taken from Cavadias (1989) show the diagrams \((V_1, V_2)\) and \((W_1, W_2)\) of the application of this method to the basins of the Island of Newfoundland. In these figures, the basin G was not included in the initial computation and its location in the two figures was computed on the basis of its physiographic characteristics alone.

In light of the results of the canonical correlation method, it appears that the problem of determining hydrologically homogeneous regions should be turned upside down, i.e. instead of seeking a pre-determined region in

\[ \text{Fig. 4} \]
which to classify a given ungauged basin, one should seek the optimal
neighbourhood of a given basin in the space of the flood-related variables for
estimating the flood distribution.

If the basin variables are good predictors of the flood quantities we can
examine the characteristics of the basins in the \( W_1, W_2 \) neighbourhood of the
ungauged basin in order to estimate the characteristics of its maximum annual
floods.

We shall use this procedure for basin G which was omitted from the
computation in the previously mentioned study. An examination of Fig. 4
shows that the six basins closest to G are F, Q, K, O, E, A in that order.
These basins constitute a "neighbourhood" of basin G. The value of the
relevant basin variables (latitude, longitude, drainage area, percentage area
controlled by lakes and swamps, mean annual runoff and shape factor) for
these basins are shown in Table 1. An examination of this table shows that
the ranges of all variables of the basins in the neighbourhood of G are
smaller than the overall ranges i.e. these basins can be considered as "similar"
to basin G with respect to basin variables related to flood characteristics. It
follows that the mean values of the variables \( Q_2/A \) and \( Q_{100}/Q_2 \) for the
neighbourhood basins can be used as a rough indication of their
magnitudes for basin G. If the number of basins in the neighbourhood is
sufficiently large, their basin variables can be used in a multiple regression
study to estimate the flood characteristics of the ungauged basin. Other sizes
of neighbourhood must also be considered in order to decide on the number
of basins and the degree of similarity of the basins in the neighbourhood with
the ungauged basin.

CONCLUSION

The results of the study show that the \( a \ priorti \) determination of
"homogeneous" regions and the use of pre-determined regression equations for estimating the flood distributions of ungauged basins does not do justice to the difficulty of the flood estimation problem.

The proposed approach of determining the "neighbourhoods" of an ungauged basin in the spaces of the canonical variables which are determined on the basis of flood characteristics is a step toward a more flexible estimation of the distribution of maximum floods and enables the design engineer to make greater use of this knowledge of the basin for assessing the similarity of the basin with the other basins in its neighbourhood.

The development of criteria regarding the optimal number of basins in a neighbourhood requires further study. These criteria should lead to a compromise between the conflicting objectives of the degree of similarity between the basins and the sample sized used for the estimation of regression coefficients.

REFERENCES


