Record Number:	21190
Author, Monographic:	Cavadias, G. S.//Ouarda, T. B. M. J.//Bobée, B.//Girard, C.
Author Role:	
Title, Monographic:	A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins
Translated Title:	
Reprint Status:	
Edition:	
Author, Subsidiary:	
Author Role:	
Place of Publication:	Québec
Publisher Name:	INRS-Eau
Date of Publication:	2001
Original Publication Date:	Janvier 2001
Volume Identification:	
Extent of Work:	ii, 54
Packaging Method:	pages
Series Editor:	
Series Editor Role:	
Series Title:	INRS-Eau, rapport de recherche
Series Volume ID:	578
Location/URL:	
ISBN:	2-89146-446-X
Notes:	Rapport annuel 2000-2001
Abstract:	Rédigé pour NSERC/Hydro-Québec
Call Number:	R000578
Keywords:	rapport/ ok/ dl

A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins.

Research Report No R-578 *January 2001*

A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins.

G.S. Cavadias

T.B.M.J. Ouarda

B. Bobée

C. Girard

NSERC/Hydro-Québec Chair in Statistical Hydrology
Institut National de la Recherche Scientifique, INRS-Eau
2800 Einstein, C.P. 7500, Sainte-Foy (Québec) G1V 4C7

Research Report No. R-578

January 2001

ISBN : 2-89146-446-X

1. ABSTRACT

This paper describes a canonical correlation method for determining the homogeneous regions used for estimating flood characteristics of ungauged basins. The method emphasizes graphical and quantitative analysis of relations between the basin and the flood variables before the data of the gauged basins are used for estimating the flood variables of the ungauged basin. The method can be used for both homogeneous regions determined a priori by clustering algorithms in the space of the flood-related canonical variables as well as for «regions of influence» or «neighbourhoods» centered on the point representing the estimated location of the ungauged basin in that space.

2. RÉSUMÉ

Détermination des régions homogènes pour l'estimation régionale de crues de bassins non jaugés

Résumé

Cet article décrit l'application de l'analyse canonique des corrélations à l'estimation régionale des crues annuelles maximales. La méthode projetée met l'accent sur l'étude des relations entre les variables de bassin et de crue des bassins jaugés avant leur utilisaiton pour l'estimation des crues de bassins non jaugés. Cette méthode peut être utilisée pour la détermination de régions homogènes obtenues par des algorithmes de classification ou des "voisinages hydrologiques" ou "régions d'influence" dont le centre est le point correspondant au bassin non jaugé.

3. INTRODUCTION

The flood characteristics of ungauged basins are estimated by regional methods i.e. by using relationships between physiographical and meteorological variables and characteristics of the maximum annual floods or the partial duration series of a set of gauged basins with hydrological regimes similar to those of the ungauged basin.

The usual steps of the estimation are:

(1) Determination of a set of similar basins («homogeneous region»).

(2) Regional estimation of the flood distribution of the ungauged basin.

Homogeneous regions may be defined in the space of geographical coordinates. This definition, however, has the disadvantage that it is not applicable to small areas and that contiguous basins may not be hydrologically similar (Linsley 1982, Cunnane 1986, Wiltshire 1986). To overcome this difficulty, some researchers have defined homogeneous regions in the space of flood-related variables (Mosley 1981, Gottschalk 1985, Wiltshire 1986). This definition has both advantages and disadvantages. The primary advantage is that it is based on variables directly related to the flood phenomenon; its disadvantages are firstly the difficulty of relating the characteristics of hydrologically defined homogeneous regions to the topographical, physiographical and meteorological conditions of the area and secondly the fact that homogeneous regions in this definition are usually determined by cluster analysis, the purpose of which is to discover «natural clusters» (Dillon and Goldstein 1984) based on the assumption that such clusters exist; however, the existence of such clusters cannot be taken for granted without prior testing (Rogers 1974, Dubes and Zeng 1987). The final set of homogeneous regions depends on the clustering method, the initial partitioning of the space and the metric used. For this reason, some researchers attempted to relate this type of homogeneous region to the geographical coordinates empirically (Mosley 1981, Gottschalk 1985), or to introduce the concept of fractional membership of a basin to a homogeneous region (Wiltshire 1986).

An entirely different concept of homogeneous region is the «neighbourhood» or «region of influence». Here a homogeneous region is defined in the space of physiographical, meteorological and hydrological variables and centered on the basin under investigation (Acreman and Wiltshire 1989, Burn 1990ab, Zrinji and Burn, 1994, Ouarda et al., 1998). This type of region avoids the difficulties related to the existence of «real» clusters but, in contrast to a priori regions, it has to be determined specifically for each basin under investigation.

According to the region of influence method, the gauged basins enter the region of influence in the order of their weighted euclidean distances from the ungauged basin in the space of the physiographical and meteorological variables where the weights of the variables are selected by the user. At every step, a homogeneity test, based on the flood distributions of the gauged basins of the homogeneous region is used to determine whether the boundary of the homogeneous region has been reached. The «region of influence» approach has the following limitations:

(a) It requires a choice of an arbitrary weight for each basin variable.

(b) It uses weighted euclidean distances that do not take into account the correlations between the basin variables.

(c) The region of influence is determined using both the basin and the flood variables without taking into account the relations between these two sets of variables.

Another approach for determining basin-centered homogeneous regions, introduced by Cavadias (1989, 1990) and Ribeiro-Correa et al (1994), uses the multivariate method of canonical correlation analysis which takes into account the relationships between the physiographical and meteorological variables and the characteristics of the distribution of the maximum annual floods. As a first approximation, linear relationships between the variables are assumed.

In a recent paper, Bates et al. (1998) classify a set of Australian basins in homogeneous regions on the basis of the L-moments of their flood characteristics. The results of this classification are verified by several multivariate techniques i.e. principal components, cluster analysis, canonical variate analysis, canonical correlation and tree based modelling of the meteorological and physiographical variables. The use of canonical correlation is restricted to a comparison of the canonical variable scores and loadings for the homogeneous regions, determined on the basis of L-moments. The paper does not address the problem of estimating the flood characteristics of an ungauged basin. More specifically, the authors do not examine the adequacy of the meteorological and physiographical basin variables for estimating the flood distribution of the ungauged basin, nor do they propose a method for classifying such a basin in one of the homogeneous regions.

The purpose of the present paper is to describe the use of canonical correlation for determining the basin-centered homogeneous region or neighbourhood of an ungauged basin. The proposed method is applied to the basins of the province of Ontario, one of which is considered ungauged. The emphasis is on the development of statistical methodology. A complete study of the flood hydrology of an ungauged basin requires, in addition to the statistical analysis, a detailed investigation of the climatology, meteorology, and geomorphology of the basin and its geographical environment.

In a recent intercomparison study (GREHYS, 1996a,b) the following methods of determining homogeneous regions for flood estimation were compared for the basins of the Canadian provinces of Ontario and Quebec: Correspondence analysis, hierarchical clustering, canonical correlation and L-moments. The study concludes that "the specific use of the canonical correlation technique yielded the best results for the delineation of homogeneous regions.

4. THE CANONICAL CORRELATION METHOD

The method of canonical correlation was developed by Hotelling (1935) and introduced into hydrology by Torranin (1972) but, despite its theoretical interest, has not found many applications in data analysis mainly due to the difficulty of interpretation of the canonical variables (e.g. Kendall and Stuart 1968, vol. 3). The application of the method described in this paper emphasises the interpretation of the configurations of sample points in the spaces of uncorrelated basin and flood variables. In addition, a comparison of the scores and loadings of the canonical variables may be useful, particularly in the case of "a priori" defined homogeneous regions, as in the case of Bates et al. (1998). A comparison of a priori defined and basin-centered homogeneous regions is presented in Cavadias (1985).

The basic idea of the method is indicated in fig. 1: Starting from the data matrices X and Q of the basin-related and flood-related variables of a set of gauged basins, we compute the canonical correlation coefficients r_1, r_2, etc. and the two matrices V and W of the corresponding canonical variables. The next step is to represent the basins as points in the spaces of the pairs of uncorrelated canonical variables and examine the similarity of the point-patterns in these spaces, i.e. the capability of basin-related variables to represent flood-related variables. If the point-patterns are sufficiently similar, we proceed to identify homogeneous sub-regions in the space of the flood-related canonical variables. At this point, we have two alternatives: If there are well defined clusters, we delineate a number of fixed homogeneous sub-regions, classify the ungauged basin in one of the homogeneous sub-regions according to its coordinates in the space of the flood-related canonical variables computed from the corresponding coordinates in the space of the basin-related canonical variables and use the basins of this sub-region to estimate its flood characteristics. In the absence of clusters we use the computed point in the space of the flood-related canonical variables as a center of a hydrological neighbourhood, the basins of which are used for the estimation of the flood characteristics of the ungauged basin.

The first of the above alternatives is discussed in Cavadias (1989) and the second in Cavadias (1990), Ribeiro-Correa et al (1994) and in Ouarda et al. (2000). The present paper describes in more detail the hydrological and statistical aspects of the second alternative i.e. the delineation of the hydrological neighbourhood of an ungauged basin.

A brief outline of canonical correlation in the context of regional flood estimation is given in Appendix A.

The method is applied in three steps:

(1) Analysis of gauged basins with the purpose of determining whether the chosen basin variables provide sufficient information for the estimation of the flood characteristics of the ungauged basin.

(2) Delineation of the homogeneous region (neighbourhood) of the ungauged basin Z.

(3) Estimation of the flood characteristics of the ungauged basin Z.

This paper deals mainly with the first two steps: Although the canonical correlation method can also be used for the third step, the intercomparison of methods of flood estimation carried out by the GREHYS group (GREHYS 1996 b) showed that the canonical correlation method is the most efficient for the first two steps whereas regression methods, which are equivalent to the canonical correlation method, are less efficient than the index flood method for the third step.

In the following paragraphs we describe in more detail the computational steps of the estimation of the flood distribution quantiles of the ungauged basin.

Step 1

1.1 Selection of the geographical, physiographical and meteorological basin variables (x_1, ..., x_p) and the flood-related variables (q_1, ..., q_m) (e.g. quantiles of the distribution of maximum floods) where usually $p \geq m$. The selection is based on hydrological considerations and data availability. The selected variables should be transformed to normality.

1.2 Calculation of the canonical correlation coefficients $r_1(v_1, w_1)$, r_2 (v_2, w_2) etc. and the two sets of canonical variables (v_1, ..., v_m) and (w_1, ..., w_m).

1.3 Examination of the corresponding point-patterns in the pairs of scatter diagrams [(v_1, v_2), (w_1, w_2)], [(v_1, v_3), (w_1, w_3)] etc. for determining whether there are clusters of points or outliers and whether the corresponding point-patterns are similar (Fig. 2). The similarity of the patterns is related to the significance of the canonical correlation coefficients (If $r_1 = r_2 = ... = r_m = 1$, the patterns are identical). Bartlett's test of significance of the canonical correlation coefficients is described in the appendix A.

There have been many attempts at developing «objective» indices of similarity of two multidimensional point-patterns (Andrews and Inglehart 1979, Leutner and Borg 1983, Borg and Leutner 1985, Borg and Lingoes 1987), based on the set of distances between all pairs of points in the two configurations. It must be noted, however, that the correlation coefficient between these distances is a misleading index because even if it is equal to one, the patterns may be different [For example, if in the (v_1, v_2) diagram the distances of the points A, B, C are AB=1 BC=2 and AC=3, the three points A, B, C are on a straight line. If in the (w_1, w_2) diagram the distances of the corresponding points are AB=3, BC=4 and CA=5, the three points

A, B, C form a right triangle. However, the correlation coefficient of the pairs (1,3), (2,4), (3,5) is equal to one].

To avoid this problem, Borg and Lingoes (1987) propose to use the correlation coefficient computed on the basis of a regression line through the origin of the distance space. This "congruence coefficient" c is computed from the formula:

$$c = \frac{\sum_{i=1}^{l} d_{iv}\, d_{iw}}{\left(\sum_{i=1}^{l} d_{iv}^2 \cdot \sum_{i=1}^{l} d_{iw}^2\right)^{\frac{1}{2}}}$$

where:

d_{iv} = ith distance of two points in the space $(v_1, ..., v_m)$

d_{iw} = ith distance of two points in the space $(w_1, ..., w_m)$

$l = \frac{n(n-1)}{2} =$ number of distances between pairs of points.

The distribution of c is mathematically intractable. Leutner and Borg (1983) give graphs, based on simulations, of the 5% level of significance of this coefficient as a function of the dimension of the spaces of the scatter diagrams and the sample sizes. It must be noted, however, that the congruence coefficient attempts to condense the information of two scatter diagrams in a single number and therefore cannot show, like the scatter diagrams, where the configurations match and where they do not (Borg and Lingoes 1987). Similarly, a correlation coefficient does not provide all the information contained in a scatter diagram, even in the case of linearly related variables.

The adequacy of the basin variables for estimating the flood variables can be tested more directly in the following way: (Cavadias 1995): In addition to the values of the flood-related canonical variables (w_1, ..., w_m) computed as linear combinations of the flood variables, we estimate the canonical variables $(\hat{w}_1,\cdots,\hat{w}_m)$ using the simple regressions on the corresponding variables (v_1, ..., v_m) and for each basin B_i we plot the «error vectors» $\hat{B}B_i$ (Fig. 3). The vector $\hat{B}B_i$ represents the difference between the local (B_i) and the regional ($\hat{B}_i$) estimation of the location of the ith basin in the space (w_1, ..., w_m). A study of the lengths and directions of the error vectors for all basins B_i ($i = 1, 2, ..., n$) enables the user to discover outliers and local patterns and relate them to the causative factors of the annual floods.

Corresponding error vectors can also be computed in the space (q_1, ..., q_m) of the original flood related variables in which they can be interpreted directly (Fig. 3).

The error vectors are also useful in the case of a combination of local and regional estimates (Kuczera 1982, Bernier 1992). The combined estimate is represented by a point on the error vector that lies between the points $\hat{B}_i$ and B_i and its location depends on the respective variances of the regional and local estimates.

1.4 If the analysis of the previous paragraph provides a satisfactory estimation of the maximum floods of the gauged basins, it is useful to determine the proportions of the variances of the variables (q_1, ..., q_m) that can be explained by the basin variables (x_1, ..., x_p) and the relative importance of various sub-groups of geographical, physiographical and meteorological variables. This is accomplished by an analysis of the structure correlation matrices i.e. the matrices of the correlation coefficients of the original and canonical variables and the study of the «redundancy indices» proposed by Stewart and Love (1968). The computation of these indices is described in appendix A.

It is also important to examine the stability of the canonical correlation coefficients and the coefficients of the canonical variables. This is achieved by either subdividing the group of gauged basins into two or more sub-groups and repeating the computations for each sub-group or by using the jackknife method (e.g. Mosteller and Tukey 1977) which has the advantage of providing approximate confidence intervals for the estimated coefficients. Another advantage of the jackknife method is that each gauged basin is considered in turn as ungauged and its flood characteristics are computed using the remaining basins.

<u>Step 2</u>

2.1 Computation of the basin-related canonical variables [$v_1(Z)$, ..., v_m (Z)] of the ungauged basin Z as linear combinations of the basin variables.

2.2 Estimation of the canonical variables $[\hat{w}_1, (Z), \cdots, \hat{w}_m(Z)]$ using the simple regressions with the canonical variables v_1 (Z), ..., v_m (Z).

2.3 **(a)** Calculation of the Mahalanobis distances M(i) of each gauged basin (i) from the estimated location of the ungauged basin. The Mahalanobis distance metric is a generalisation of the Euclidean distance adjusted to take into account the correlations between the variables. Other metrics could also be used.

(b) Sorting of the variable M(i) in descending order and determining a sequence of neighbourhoods of diminishing size by eliminating in turn the most distant basin.

These neighbourhoods become progressively more homogeneous and consist of basins more similar to the ungauged basin which is at the centre of the neighbourhoods.

A simple measure of homogeneity of a neighbourhood is the standard deviation of the basin and flood variables after a normalizing transformation. More elaborate homogeneity tests have been proposed in the literature (Wiltshire 1986, Chowdhuri et al 1991).

A set of tests based on the three L-moments (L-Cv, L-Cs and L-Kurtosis) of the distributions of the maximum annual floods of each basin B_i of the neighbourhood was proposed by Hosking and Wallis (1993). These tests were used in the inter-comparison study of flood frequency procedures (GREHYS 1996b) mentioned previously.

(c) The best neighbourhood is chosen on the basis of the minimum size of the prediction intervals for the flood variables of the ungauged basin, computed from the multiple regressions of the flood variables on the basin variables for each neighbourhood.

This statistical procedure for choosing the best neighbourhood must be supplemented by an examination of all relevant physical factors to ensure that the chosen neighbourhood makes hydrological sense and is not just an artifact of the statistical calculations.

Step 3

Several methods of estimation of the flood characteristics of an ungauged basin using the data of the basins of its neighbourhood have been reviewed and compared [GREHYS 1996a, b] and it was concluded that the following three methods called REM 1, REM 2 and REM 3 respectively, give the best results for the Canadian basins considered in the study:

REM 1. Generalized extreme value / PWM index flood procedure (Dupuis and Rasmussen 1993).

REM 2. Regional non-parametric analysis (Gingras and Adamowski 1992).

REM 3. Regional flood estimation by peaks-over threshold (POT) methods (Ouarda and Ashkar 1995).

It is indicated in appendix A that the estimation of the flood variables of the ungauged basin by canonical correlation is equivalent to the estimation by linear multiple regression on the basin variables. The regression method of estimation was compared to other current methods in GREHYS (1996b) for data from Quebec and Ontario, a subset of which was used in this paper. The conclusions of that study are:

(a) The best regression model for these data is the multiplicative model of the form $Q_T = a_o \bullet x1^{k_1} \bullet\bullet\bullet\, x_p^{kp} \bullet e$

where $x_1, \ldots, x_p$ are basin characteristics and $k_1, \ldots, k_p$ are parameters estimated by nonlinear optimization.

(b) The above non-linear regression model performed less well than the three best estimation methods mentioned previously. Consequently, upon delineation of the best neighbourhood by canonical correlation, the selection of the most efficient estimation method must take into account the results of the GREHYS study or similar comparisons. Such a comparison of estimation methods is beyond the scope of this paper and therefore, in the example given in a later section, we estimate the flood variables of the ungauged basin using the linear regressions on the basin variables of the best neighbourhood.

It is to be noted that the authors of the GREHYS paper (GREHYS, 1996b) state that the results of the regional analysis are affected more by the delineation of the homogeneous regions than by the method of flood estimation and conclude that «all computational methods associated with the canonical correlation method for the identification of the neighbourhood appear to give good results.»

A different approach for the estimation of the flood characteristics of ungauged basins on the basis of a given homogeneous region was proposed by Roy (1993). According to this method, a conceptual precipitation-runoff model is calibrated for the nearest gauged basin of the neighbourhood of the ungauged basin in the space of the canonical variables (w_1, ..., w_m) and used to simulate the daily discharges of the ungauged basin on the basis of its known meteorological inputs. The next step is to fit a probability distribution to the simulated maximum daily discharges for each year and estimate the floods of the required return periods. An important limitation of the method is that it deals with maximum daily and not instantaneous peaks. However, it is a useful complementary approach to the usual procedures needing further development.

5. APPLICATION

The canonical correlation method is applied to the estimation of the maximum annual floods of the Province of Ontario in Canada. The locations of the 106 basins are shown in figure 4. We consider basin (46) as ungauged and use the remaining 105 basins to estimate its flood variables.

Step 1

The following basin and flood variables are used:

Basin variables

LUAIRE = $\log_{10}$ [Drainage area (km^2)].

LUPCP = $\log_{10}$ [Slope of main river channel (m/km)].

LULCP = $\log_{10}$ [Length of main river channel (km)].

LUSLM = $\log_{10}$ (SLM + 0.01) where SLM= Area of drainage basin controlled by lakes and swamps (km^2).

LUPTMA = $\log_{10}$ [Mean annual precipitation (mm)].

Flood variables

LUQ2 = $\log_{10}$ [Annual maximum flood of 2-year return period (m^3/sec)].

LUQ1002 = $\log_{10}$ [Ratio of the 100-year maximum flood to the 2-year maximum flood).

The Kolmogorov-Smirnov test was used to examine the normality of the transformed variables. The results showed that the normality assumption cannot be rejected, except for the variable LUSLM which has many equal values corresponding to SLM=0.

We form the (105x5) matrix of the basin variables and the (105x2) - matrix of the flood variables and carry out the canonical correlation computations based on the 105 gauged basins. The resulting canonical correlation coefficients $r_1 = 0.96$ and $r_2 = 0.33$ are both significant at the 5% level according to Bartlett's test.

The scatter diagrams of the two pairs of canonical variables (v_1, v_2) and (w_1, w_2) are shown in figures 5 and 6 respectively. An examination of these diagrams shows that:

(a) The locations of points corresponding to the same basin in the two diagrams are approximately similar.

(b) These are no clearly defined clusters of points and consequently the basin-centered homogeneous region (neighbourhood) approach is more appropriate for this problem.

An examination of the structure correlation matrices (Table 1) shows that, as expected, the correlations of the canonical variables v_1 and w_1 with the basin and flood variables are higher than those of the canonical variables v_2 and w_2. The squares of the structure correlation coefficients are used for the computation of the redundancy indices. The overall index $R_{q/v} = 0.548$ is the sum of the contributions of the two sets of canonical variables which are respectively 0.497 and 0.051.

The diagrams of error vectors using the initial number of 105 basins are not shown because the large number of points makes the interpretation difficult.

The estimated coordinates of the point corresponding to the ungauged basin are $\hat{w}_1(46) = -0.19$ and $\hat{w}_2(46) = -0.27$. We proceed to compute the Mahalanobis distances M(i) of each gauged basin (i) from this point and to arrange them in descending order (Fig. 7). These distances are used to form a sequence of neighbourhoods of diminishing sizes by consecutive omission of the basin with maximum M(i). These neighbourhoods have increasing homogeneity as indicated by the decreasing standard deviations of the flood variables LUQ2 and LUQ1002 (Figure 8). The estimated values of these variables as well as the corresponding 95% prediction intervals are plotted against the size of the neighbourhood in figures 9 and 10 which indicate that the minimum value of these intervals corresponds to a neighbourhood of 20 basins, which is then used for the estimation of the flood variables of the ungauged basin (46). The multiple correlation coefficients for neighbourhoods with 17 or fewer basins are not significant.

The canonical correlations for the 20-basin neighbourhood are $r_1 = 0.87$ and $r_2 = 0.65$ and their significance levels are respectively 0.001 and 0.1. The structure correlation matrices for 20 basins are shown in Table 2. The overall redundancy index is $R_{q/v} = 0.555$ i.e. slightly higher than the redundancy index corresponding to 105 basins. Since $x^2 = 2.53$ for the most distant basin, the 20-basin neighbourhood corresponds to a 63% confidence region. The F-test which is appropriate in the case of estimated covariance matrix results in a confidence region of 65%.

Tables 3 and 4 present respectively a comparison of the means and standard deviations of all basin and flood variables and the results of the multiple regressions of the flood - on the basin variables, for the neighbourhoods of 105 and 20 basins. An examination of these statistics shows that the standard deviations of all basin and flood variables (except for the basin variable (PTMA) are smaller for the 20-basin neighbourhood than for the initial 105 basin neighbourhood. Although the adjusted squared correlation coefficient $\overline{R}^2$ of the basin variable LUQ2 is smaller for 20 basins than for 105 basins, the corresponding prediction interval is smaller because of the smaller standard deviation of LUQ2 for this neighbourhood.

Thus, the 20-basin neighbourhood is more satisfactory for estimating the flood variables of the ungauged basin (46) because it is more homogeneous.

Figures 12 and 13 show the error vectors for the 20-basin neighbourhood in the spaces (w_1, w_2) and $(Q_2, Q1002)$ respectively. The latter diagram whose axes are in the original units of the flood variables gives an intuitive picture of the performance of the canonical correlation method.

These two diagrams must be studied in detail to discover the reasons for the different lengths and directions of the error vectors for the different gauged basins of the neighbourhood.

Figure 11 shows the geographical locations of the basins of the 20-basin neighbourhood. The fact that four basins of north-western Ontario are grouped with basins of south-western Ontario requires further hydrological analysis.

6. CONCLUDING REMARKS

The most important features of the canonical correlation method are:

(1) Before proceeding to the estimation of the flood characteristics of the ungauged basin, it includes a detailed study of the relationships between the basin and the flood variables of the gauged basins to ensure that the latter variables can be estimated from the former.

(2) The homogeneous region used for the estimation is determined in the canonical space of the flood variables which is based on the relationships between the basin and the flood variables.

The method of canonical correlation was compared to other current methods of delineation of homogeneous regions such as regions of influence (Zrinji and Burn 1994), correspondence analysis (Birikundavyi et al 1993), and L-moments (Gingras et al 1994), and was found to give better results for the basins considered in the study (GREHYS 1996b)].

7. APPENDIX A

Given n basins, p standardized basin-related variables x_j and m standardized flood-related variables q_j (e.g. quantiles of a fitted probability distribution), where usually $p \geq m$, we compute the m canonical correlation coefficients $r_1 \geq r_2, \ldots, \geq r_m$ and the m pairs of standardized basin - and flood-related canonical variables v_j and w_j respectively.

The flood-related canonical variables w_j can be estimated from the corresponding basin-related canonical variables v_j using the simple regression equations:

$$\hat{w}_j = r_j v_j (j = 1, 2, \cdots, m)$$

It must be noted that the flood variables $(q_1, \ldots, q_m)$ can be estimated using the multiple regressions $\hat{q}_j = f_j(x_1, \cdots, x_p)$ or equivalently the multiple regressions $\hat{q}_j = f_j(v_1, \cdots, v_m)$. Thus the use of canonical variables results in a reduction of the dimensionality of the space of basin variables from p to m in a way that takes into account their relations with the flood variables. As a result, the number of flood variables that can be estimated is equal to the number of significant canonical correlation coefficients.

Statistical packages usually plot diagrams of (v_1, w_1), (v_2, w_2) etc., i.e. the pairs of canonical variables having maximum correlation coefficients. Given the difficulties in interpreting the canonical variables (e.g. Kendall and Stuart, 1968), it is preferable to plot the uncorrelated pairs of canonical variables (v_1, v_2), (v_1, v_3) ... (v_j, v_k) etc., where $j \neq k$ along with the corresponding scatter diagrams (w_1, w_2) ... (w_j, w_k) of uncorrelated flood-related canonical variables. The pairs of canonical variables (v_1, v_2) (w_1, w_2) etc. respectively define the spaces of linearly transformed basin- and flood-related variables in which the points represent individual basins. If the basin variables are good predictors of the flood-related variables, the patterns of points in the corresponding scatter diagrams are similar.

It is useful to compute the structure correlation coefficients i.e. the correlation coefficients between the original and the canonical variables which help to determine the contribution of each of the basin variables to the flood variables.

For large samples, the statistical significance of the set of m canonical correlation coefficients can be tested by Bartlett's statistic:

$$v = -(n-1-(p+1+m)/2)\sum_{j=1}^{m}\ln\left(1-r_j^2\right)$$

which, under the normality assumption, is distributed as chi-square with (pxm) degrees of freedom. If successive pairs of canonical variables are to be tested, the degrees of freedom must be modified. For example, the degrees of freedom associated with the second pair of canonical variables are (p-1) • (m-1) and so on.

The square of the canonical correlation, coefficient r_j^2 is a measure of the shared variance between the canonical variance v_j and w_j. Our main interest, however, is the proportion of the variance of the flood variables accounted for by the basin variables. An appropriate measure of this proportion is the redundancy index $Rd_{q/v}$ which is equal to the mean of the proportions of the variances of the flood variables (q_1,...q_m) accounted for by the canonical variables (v_1,..., v_m) or equivalently by the basin variables (x_1,..., x_p). The redundancy index is given by the following equations (Cooley and Lohnes 1971 p. 173):

$$Rd_{q/v} = \sum_{k=1}^{m} Rd_{q/v_k} = \sum_{k=1}^{m} Rd_{q_j/v}$$

where:

Rd_{q/v_k} = contribution of the k^th pair of canonical variables to the redundancy index $Rd_{q/v}$

$$= \sum_{j=1}^{m} \frac{r^2(q_i, v_k)}{m} = r_k^2 \sum_{j=1}^{m} \frac{r^2(q_i, w_k)}{m}$$

$Rd_{q_j/v}$ = contribution of the j^th basin variable to the redundancy index $Rd_{q/v}$

$$= \sum_{k=1}^{m} \frac{r^2(q_j, v_k)}{m}$$

Given the estimated point $\underline{\hat{w}(Z)} = \hat{w}_1(Z) \ldots \hat{w}_m(Z)$ in the m-dimensional space of the canonical variables and under the normality assumption, the (1-α) per cent confidence region for the point is given by the equation:

$M(i) = M[\underline{w(i)} - \underline{\hat{w}(Z)}] = [\underline{w(i)} - \underline{\hat{w}(Z)}]' (I_m - \Lambda)^{-1} [\underline{w(i)} - \underline{\hat{w}(Z)}] \le x^2(\alpha, m)$

where Λ is the (m x m) diagonal matrix of the squared canonical correlation coefficients $(r_1^2, \ldots, r_m^2)$ and $M\lfloor \underline{w(i)} - \underline{\hat{w}(Z)} \rfloor$ is the Mahalanobis distance of the point $\underline{w(i)}$ from the estimated point $\underline{\hat{w}(Z)}$.

This confidence region can be interpreted as the $(1-\alpha)$ per cent neighbourhood of the point $\underline{\hat{w}(Z)}$ (Ribeiro-Correa et al 1994). In the special case of m=2 the above equation is simplified:

$$M[\underline{w(i)}, \underline{\hat{w}(Z)}] = \frac{[w_1(i) - w_1(Z)}{1 - r_1^2} + \frac{[w_2(i) - w_2(Z)}{1 - r_2^2} \le x^2(\alpha, 2)$$

If the normality assumption is not valid, the Mahalanobis distance is a weighted distance of the basin $\underline{w(i)}$ from the estimated location $\underline{\hat{w}(Z)}$ of the ungauged basin, with weights depending on the canonical correlation coefficients.

8. APPENDIX B : NOTATION

α	Significance level.
$\mathbf{k_1, k_p}$	Regression parameters.
$\mathbf{d_{iv}}$	i^{th} distance of two points in the space $(v_1,\ldots, v_m)$.
$\mathbf{d_{iw}}$	i^{th} distance of two points in the space $(w_1,\ldots, w_m)$.
e	Regression error.
LUAIRE	$\log_{10}$ [Drainage area (km^2)].
LULCP	$\log_{10}$ [Length of main river channel (km)].
LUPCP	$\log_{10}$ [Slope of main river channel (m/km)].
LUSLM	$\log_{10}$ (SLM + 0.01) where SLM = area of drainage basin controlled by lakes and swamps (km^2)
LUPTMA	$\log_{10}$ [Mean annual precipitation (mm)].
LUQ2	$\log_{10}$ [Annual maximum flood of 2-year return period (m^3/sec)].
LUQ1002	$\log_{10}$ [Ratio of the 100-year maximum flood to the 2-year maximum flood).
M(i)	Mahalanobis distance of gauged basin (i) from the estimated point $[\underline{w(i)}, \underline{\hat{w}(Z)}]$ of the ungauged basin Z.
m	Number of flood variables.
n	Number of basins.
p	Number of basin variables.
$\mathbf{q_1,\ldots, q_m}$	Flood variables.
$\mathbf{r_1,\ldots, r_m}$	Canonical correlation coefficients.
$\overline{R}$	Adjusted multiple correlation coefficient.

$\mathbf{Rd}_{q/v}$	Redundancy index.
$\mathbf{v}_1,\ldots, \mathbf{v}_m$	Canonical variables of the basin variables.
$\mathbf{V}$	Bartlett's statistic.
$\mathbf{w}_1,\ldots, \mathbf{w}_m$	Canonical variables of the flood variables.
$\mathbf{x}_1,\ldots, \mathbf{x}_p$	Basin variables.

9. REFERENCES

Acreman, M.C. and Wiltshire, S.E. (1989). The regions are dead: long live the regions. Methods of identifying and dispensing with regions for flood frequency analysis. IAHS Publ. no. 187, 175-1988.

Andrews, F.M. and Inglehart, R.F. (1979). The structure of subjective well being in nine Western Societies. Social Indicators Research 6, 73-90.

Bates, B.C., Rahman, A., Mein, R.G. and Weinmann P.E. (1998). Climatic and physical factors that influence the homogeneity of regional floods in southeastern Australia, Water Resources Research 34 (12) 3369-3381.

Bernier, J. (1992). Modèle regional à deux niveaux d'aléas. Interim Report NSERC Strategic Grant No STR 0118482, 11 p.p.

Birikundavyi, S., Rousselle, J. and Nguyen, V.T.B. (1993). Determination des régions homogènes pour le Québec et l'Ontario: Une approche par l'analyse des corréspondances et la classification ascendante hierarchique. Rapport final, Subvention stratégique CRSNG, No STR 0118482, École Polytechnique de Montréal, p. 44.

Borg, I. and Lingoes, D. (1987). Multidimensional similarity structure analysis. Springer-Verlag.

Burn, D.H. (1990a). An appraisal of the 'region of influence' approach to flood frequency analysis. Hydrological Sciences, Journal, 35 (2) 149-165.

Burn, D.H. (1990b). Evaluation of regional flood frequency analysis with a region of influence approach. Water Resources Research 26 (10) 2257-2265.

Cavadias, G.S. (1989). Regional flood estimation by canonical correlation. Paper presented to the 1989 Annual Conference of the Canadian Society of Civil Engineering, St. John's Newfoundland.

Cavadias, G.S. (1990). The canonical correlation approach to regional flood estimation. Regionalization in Hydrology. Proc. of the Ljubljana Symposium, IAHS. Publ. No. 191:171-178.

Cavadias, G.S. (1995). Regionalization and multivariate analysis: The canonical correlation approach. In "Statistical and Bayesian methods in hydrological sciences". IHP-V, Technical Documents in Hydrology No. 20, UNESCO, Paris.

Chowdhury, J.V., Stedinger J.R. and Lu, L.H. (1991). Goodness of fit test for regional generalized extreme value distributions. Water Resour. Res. 27(7), 1765-1776.

Cooley, W.W. and Lohnes, P.R. Multivariate data analysis. Wiley 1971.

Cunnane, C. (1986). Review of statistical models for flood frequency estimation. Keynote paper in: International Symposium on Flood Frequency and Risk Analysis (Baton Rouge, May 1986). Reidel.

Dalrymple, T. (1960). Flood frequency analysis. USGS Water Supply Paper 1534 A., p. 60.

Dillon, W.E. and Goldstein, M. (1984). Multivariate Analysis, p. 139. John Wiley.

Dubes, R. and Zeng, G. (1987). A test for spatial homogeneity in cluster analysis. Classification 4, 33-56.

Dupuis, L. and Rasmussen, R.F. (1993). Évaluation de méthodes «indice de crue» pour l'estimation d'une distribution régionale. Rapport interne INRS-Eau, No 1-125, pp. 26.

Gingras D., Adamowski, K. and Pilon, P.J. (1994). Regional flood equations for the Provinces of Ontario and Quebec. Water Resour. Bull. 30(1), 55-67.

Gingras, D. and Adamowski, K. (1992). Coupling of nonparametric frequency and L-moment analysis for mixed distribution identification. Water Resour. Bulletin: 28(2), 263-272.

Gottschalk, L. (1985). Hydrological regionalization in Sweden. Hydrol. Sci. J. (30) (1).

Green, P.E. (1978). Analyzing multivariate data. The Dryden Press, Hinsdale, Illinois.

GREHYS (1996a). Presentation and review of some methods of regional flood frequency analysis. J. Hydrol. 186 (1-4), 63-84.

GREHYS (1996b). Intercomparison of flood frequency procedures for Canadian rivers. J. Hydrol. 186 (1-4), 85-103.

Hosking, J.R.M. and Wallis, J.R. (1993). Some statistics useful in regional frequency analysis. Water Resour. Res. 29(2), 271-281.

Hotelling, H. (1936). Relations between two sets of variates. Biometrica 28:321-377.

Kendall, M.G. and Stuart, A. (1968). The advanced Theory of Statistics, Vol 3. 2nd ed. Charles Griffin & Co. London.

Kuczera, G. (1982). Combining site-specific and regional information: an empirical Bayes' approach. Water Resour. Res. Vol. 8, No. 2, pp. 306-314.

Leutner, D. and Borg, I. (1983). Zufallskritische Beurteilung der Übereinstimmung von Factor und MDS-Konfigurationen. Diagnostica 29, 320-335.

Leutner, D. and Borg, I. (1985). Zur Messung der Übereinstimmung von multidimensionalen Konfigurationen mit Indizes. Zeischrift für Sozialpsychologie, 16, 29-35.

Linsley, R.K. (1982). Flood estimates. How good are they? Wat. Resour. Res. 22 (9).

Mosley, M.P. (1981). Delimitation of New Zealand hydrological regions. J. Hydrol. 49, 173-192.

Mosteller, F. and Tukey, J.W. (1977). Data Analysis and Regression. Addison-Wesley.

Ouarda, T.B.M.J., and Ashkar, F. (1995). The peaks-over-threshold (POT) method for regional flood frequency estimation. Proceedings of the 48th Annual Conference of the Canadian Water Resources Association, Fredericton, N.B., Canada, 641-659.

Ouarda, T.B.M.J., Haché, M. and Bobée, B. (1998). Regional estimation of extreme hydrological events, INRS-Eau, Research Report No. R-534 (In French), Quebec, Canada, p. 181.

Ouarda, T.B.M.J., Haché, M., Bruneau, P. and Bobée, B. (2000). Regional flood peak and volume estimation in Northern Canadian Basin, Journal of Cold Regions Engineering of the ASCE, 14(4): 176-191.

Panu, V.S., Smith, D.A. and Ambler, D.C. (1984). Regional flood frequency Analysis for the island of Newfoundland. Environment Canada and Department of Environment, Province of Newfoundland.

Ribeiro-Correa, B., Cavadias, G.S., Clement, B. and Rouselle, J. (1994). Identification of hydrological neighborhoods using canonical correlation analysis. Journal of Hydrology 173 (1995) 71-89.

Rogers, A. (1974). Statistical Analysis of Spatial Dispersion. Pion Ltd.

Roy, R. (1993). Regionalisation de Caractéristiques de Crue, Utilisation d'une Méthode Combinant les Approches Déterministes et Stochastiques, Ph.D. Thesis, INRS-Eau.

Stewart, D.K. and Love, W.A. (1968). A general canonical correlation index. Psychological Bulletin 70 (1968) 160-163.

Torranin, P. (1972). Applicability of canonical correlation in hydrology. H.P. 58, Colorado State University, p. 30.

United States Water Resources Council (1977). Guidelines for Determining Flow Frequency. USWRC, 2120 Long Island NW, Washington, DC.

Wiltshire, S.E. (1986). Regional flood analysis II: multivariate classification of drainage basins in Britain. Hydrol. Sci. J. 31 (3).

Zrinji, Z. and Burn, D.H. (1994). Flood frequency analysis for ungauged sites using a region of influence approach. Journal of Hydrology 153, 1-21.

10. TABLES

List of tables

	V1	V2	W1	W2
LUAIRE	.9468 (105) .0000	.0702 (105) .4767	.9069 (105) .0000	.0235 (105) .8120
LUPCP	-.6049 (105) .0000	.1822 (105) .0629	-.5794 (105) .0000	.0609 (105) .5368
LULCP	.9188 (105) .0000	-.1477 (105) .1326	.8801 (105) .0000	-.0494 (105) .6166
LUSLM	.4482 (105) .0000	-.2652 (105) .0062	.4293 (105) .0000	-.0887 (105) .3681
LUPTMA	-.3417 (105) .0004	-.5216 (105) .0000	-.3273 (105) .0007	-.1745 (105) .0751
LUQ2	.9549 (105) .0000	-.0264 (105) .7895	.9969 (105) .0000	-.0788 (105) .4242
LUQ1002	-.2863 (105) .0031	.3192 (105) .0009	-.2989 (105) .0019	.9543 (105) .0000

TABLE 1

	V1	V2	W1	W2
LUAIRE	.2355 (20) .3175	.0151 (20) .9496	.2049 (20) .3861	.0096 (20) .9678
LUPCP	-.0486 (20) .8388	.0734 (20) .7586	-.0423 (20) .8596	.0468 (20) .8446
LULCP	.1268 (20) .5944	.4252 (20) .0616	.1103 (20) .6435	.2715 (20) .2470
LUSLM	-.5520 (20) .0116	.4394 (20) .0525	-.4803 (20) .0321	.2806 (20) .2308
LUPTMA	.2021 (20) .3927	.4849 (20) .0302	.1759 (20) .4582	.3096 (20) .1840
LUQ2	.7894 (20) .0000	.2686 (20) .2522	.9072 (20) .0000	.4206 (20) .0648
LUQ1002	.1289 (20) .5879	-.6314 (20) .0028	.1482 (20) .5329	-.9890 (20) .0000

TABLE 2

TABLE 3

VARIABLES	OBSERVED VALUE	MEAN		STANDARD DEVIATION	
	BASIN (46)	105 BASINS	20 BASINS	105 BASINS	20 BASINS
LUAIRE	2.467	2.759	2.573	0.769	0.367
LUPCP	-0.187	0.004	-0.039	0.511	0.449
LULCP	1.903	1.758	1.633	0.430	0.171
LUSLM	-2.000	0.759	0.559	2.353	2.208
LUPTMA	2.923	2.919	2.936	0.072	0.079
LUQ2	1.777	1.758	1.617	0.556	0.149
LUQ1002	0.362	0.422	0.401	0.126	0.083

TABLE 4

MULTIPLE REGRESSION RESULTS						
DEPENDENT VARIABLES	ADJUSTED R^2		ESTIMATED VALUE		95% PREDICTION INTERVAL	
	105 BASINS	20 BASINS	105 BASINS	20 BASINS	105 BASINS	20 BASINS
LUQ2	0.908	0.586	1.664	1.651	0.672	0.418
LUQ1002	0.143	0.206	0.397	0.349	0.464	0.324

11. FIGURES

List of figures

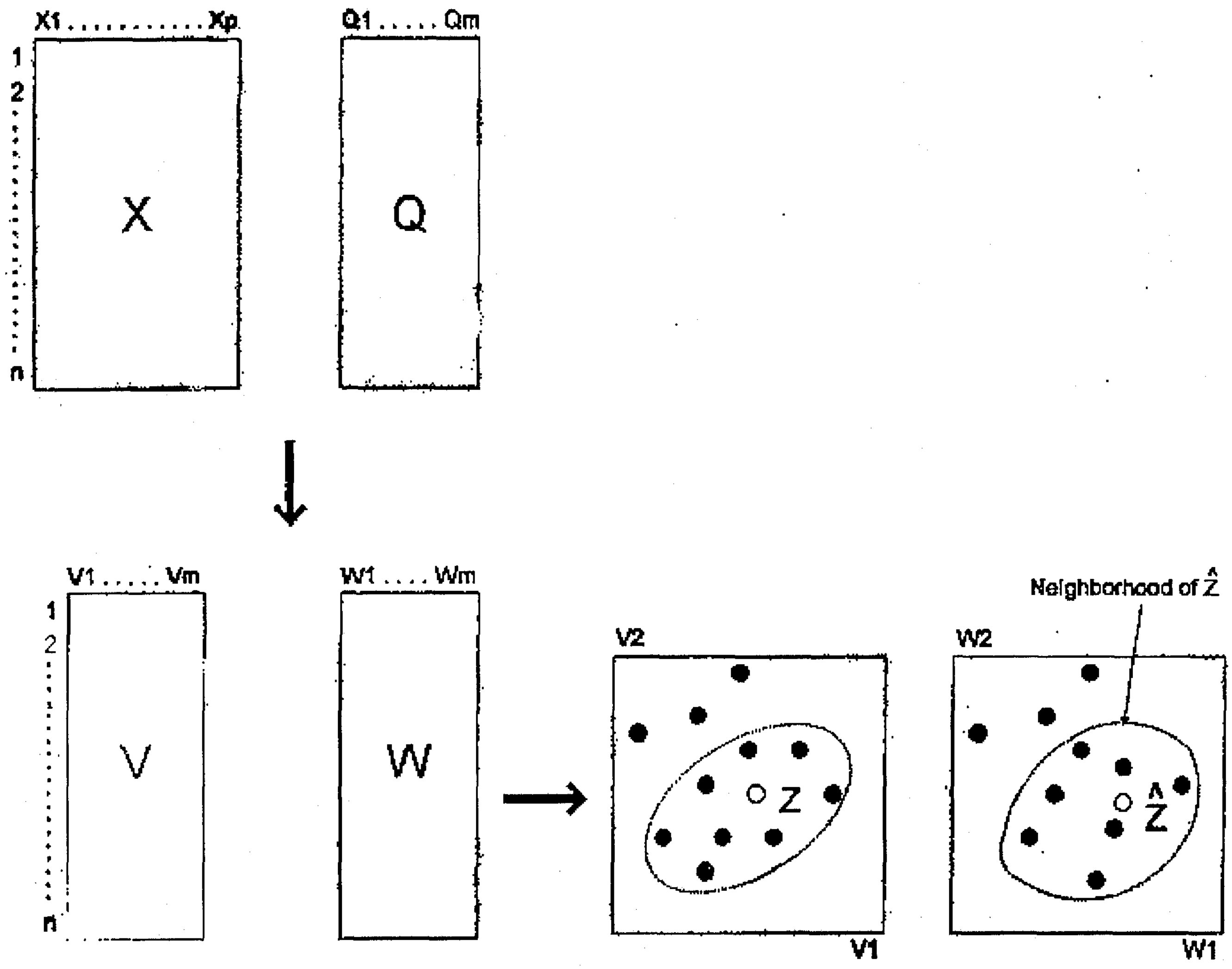
X1 . . . . Xp
Q1 . . . . . Qm
1
2
n
X
Q
V1 . . . . . Vm
W1 . . . . Wm
1
2
n
V
W
V2
Z
V1
W2
Neighborhood of $\hat{Z}$
$\hat{Z}$
W1

FIG. 1

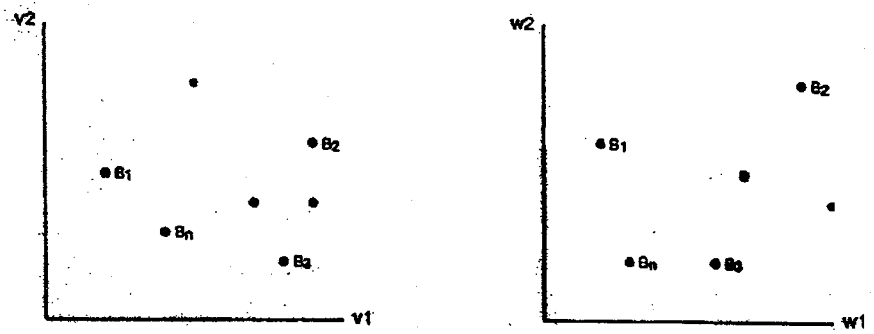
v2
B_1
B_2
B_n
B_3
v1
w2
B_2
B_1
B_n
B_3
w1

FIG 2.

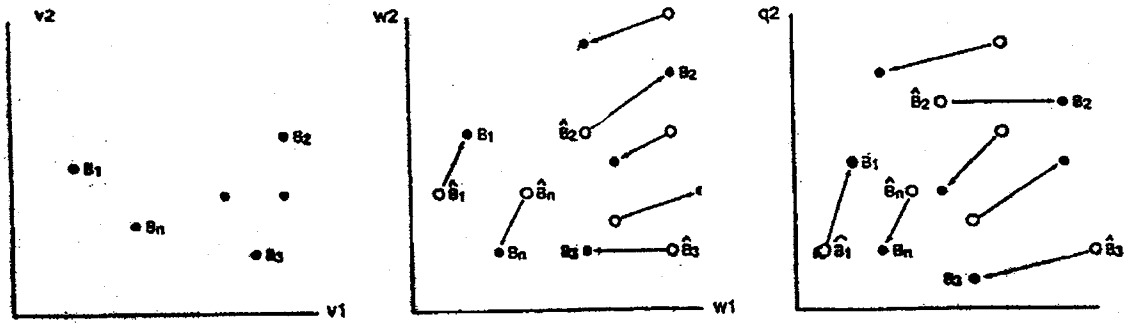
v2
B_2
B_1
B_n
B_3
v1
w2
B_2
B_1
$\hat{B}_2$
$\hat{B}_1$
$\hat{B}_n$
B_n
B_3
$\hat{B}_3$
w1
q2
$\hat{B}_2$
B_2
B_1
$\hat{B}_n$
$\hat{B}_1$
B_n
$\hat{B}_3$
B_3

FIG. 3

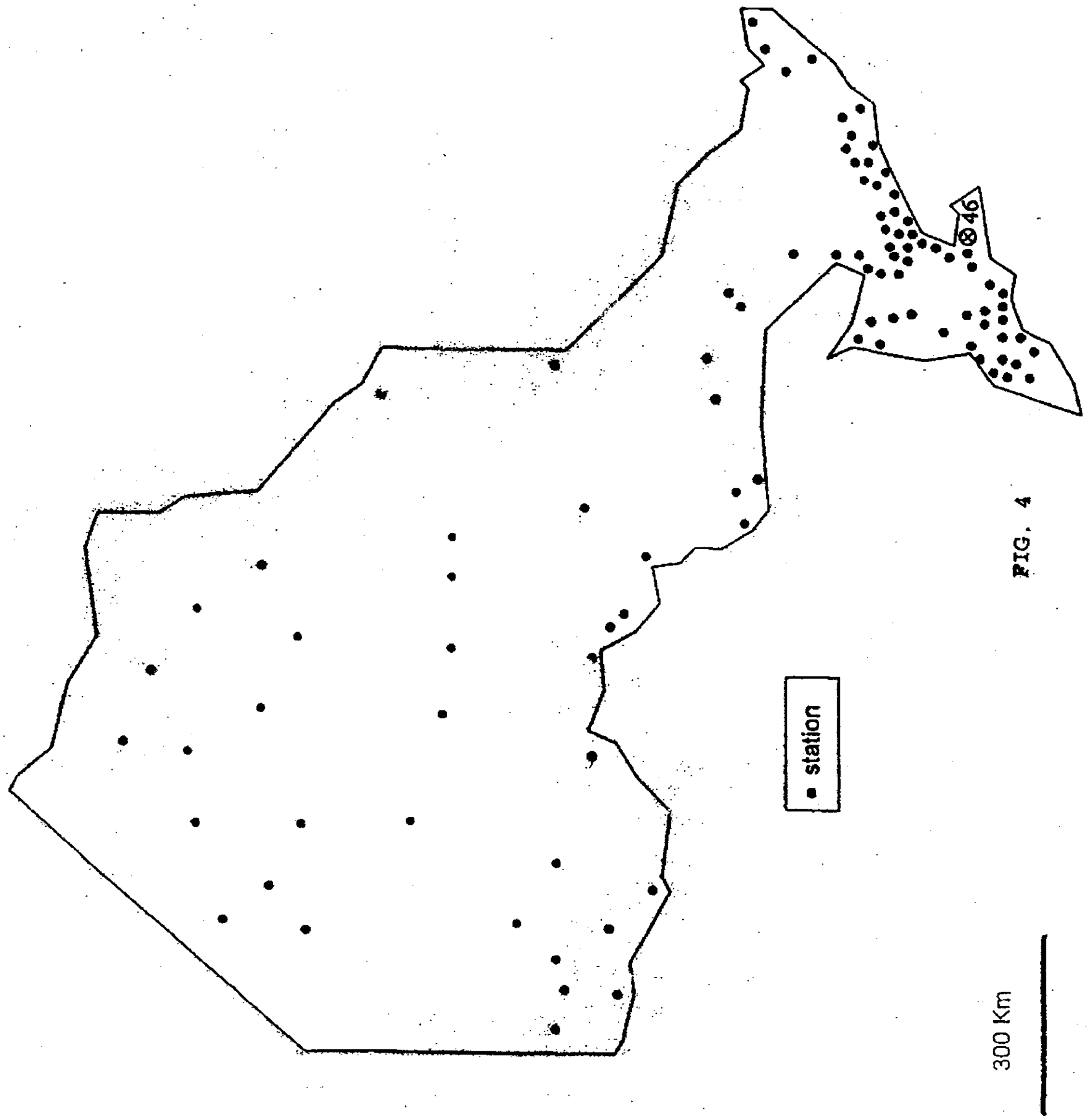

FIG. 4

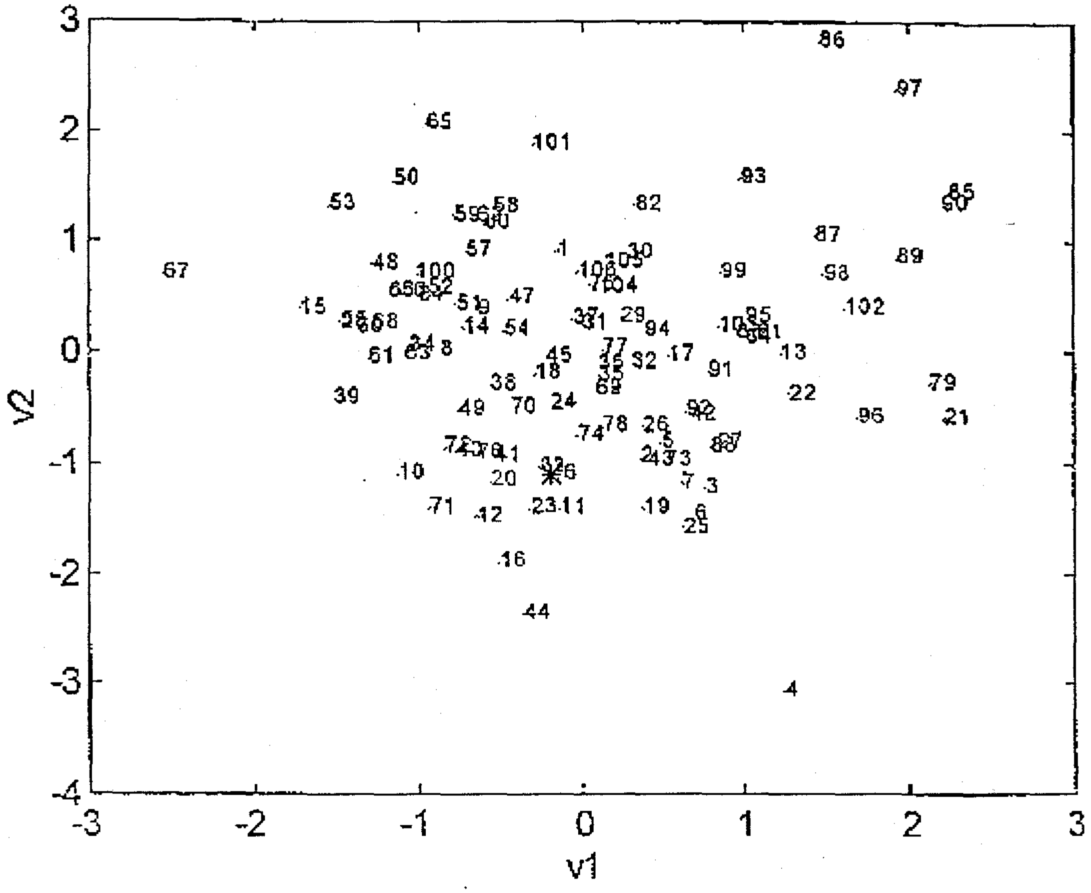
3
2
1
0
-1
-2
-3
-4
v2
-3
-2
-1
0
1
2
3
v1
86
97
65
101
50
93
53
82
58
85
90
87
57
1
30
48
100
105
89
67
106
99
98
52
104
15
51
47
102
37
31
29
94
54
14
68
8
45
77
17
13
61
92
18
91
79
38
35
39
69
22
49
70
24
96
21
74
78
26
5
10
20
73
3
7
71
12
23
11
19
6
25
16
44
4

FIG. 5

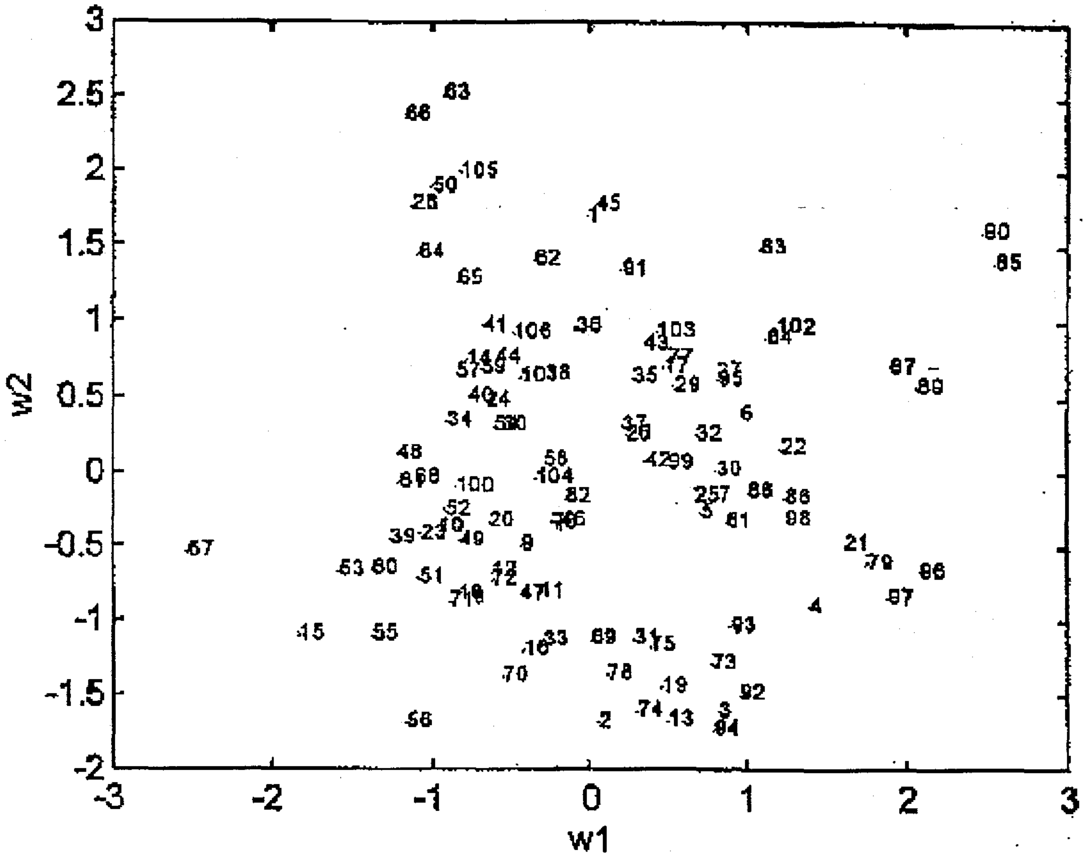

FIG. 6

PLOT OF MAHALANOBIS DISTANCE VS ORDER OF BASIN

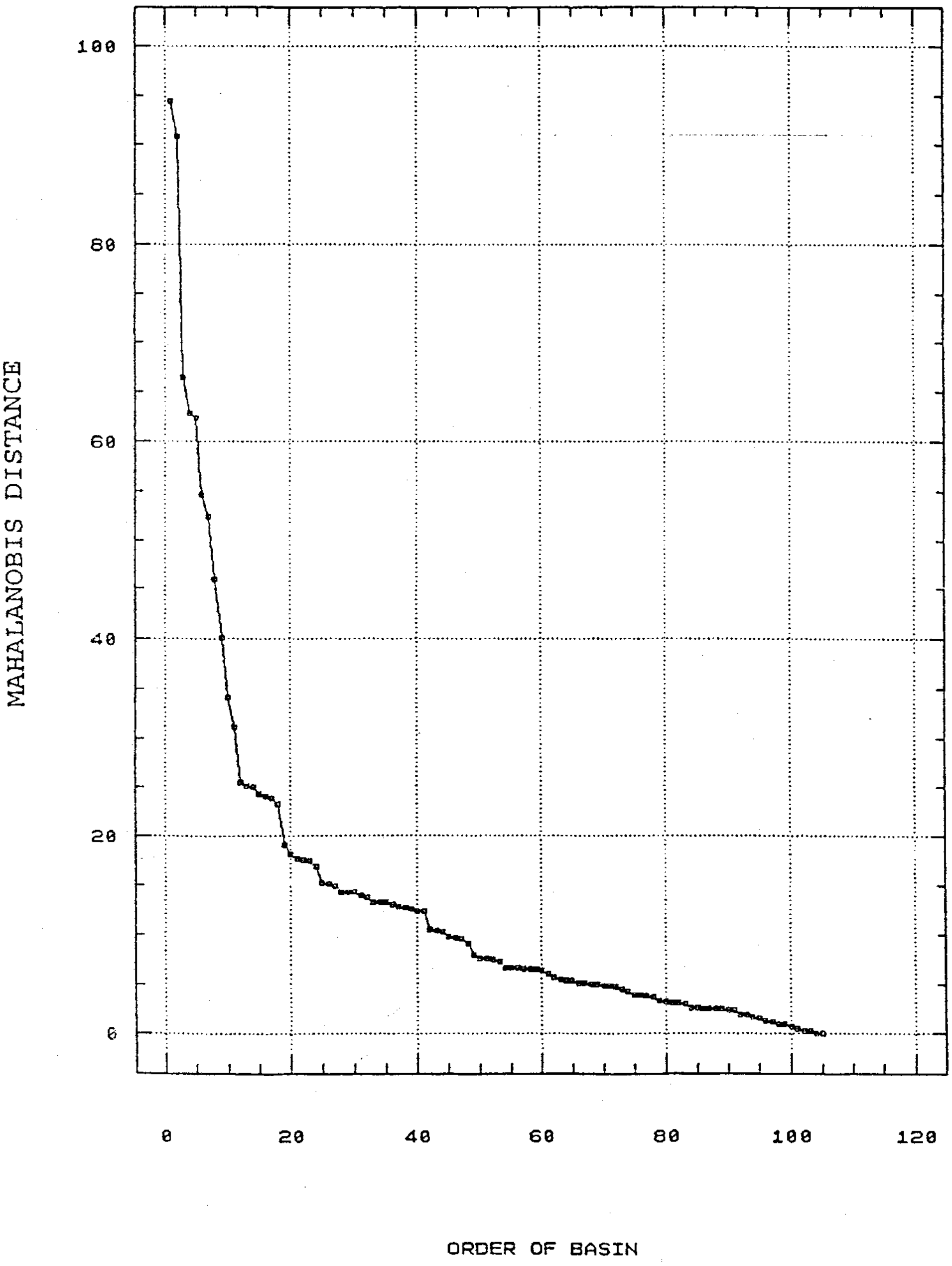

FIG. 7

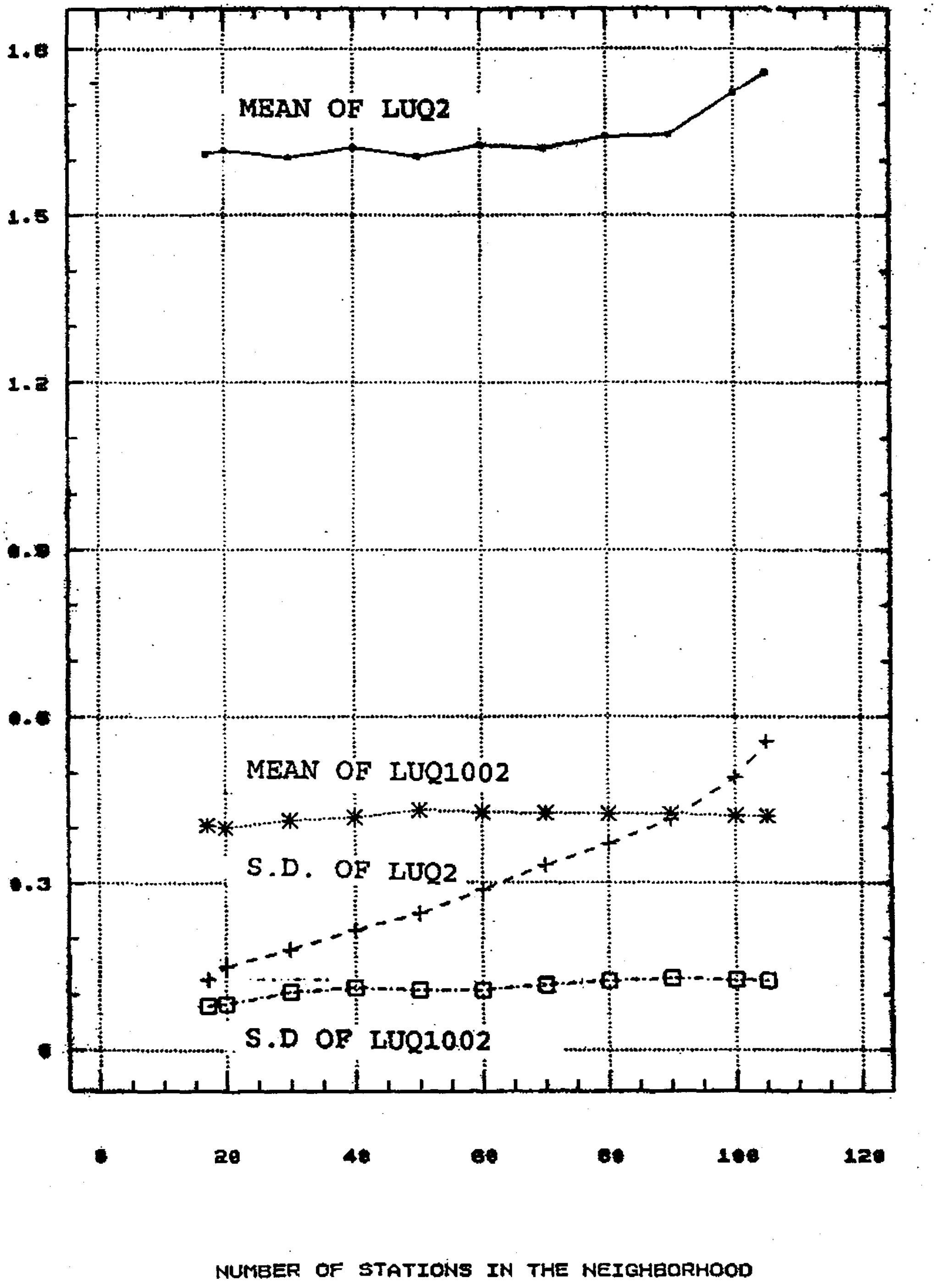
1.8
1.5
1.2
0.9
0.6
0.3
0
MEAN OF LUQ2
MEAN OF LUQ1002
S.D. OF LUQ2
S.D OF LUQ1002
0
20
40
60
80
100
120
NUMBER OF STATIONS IN THE NEIGHBORHOOD

FIG. 8

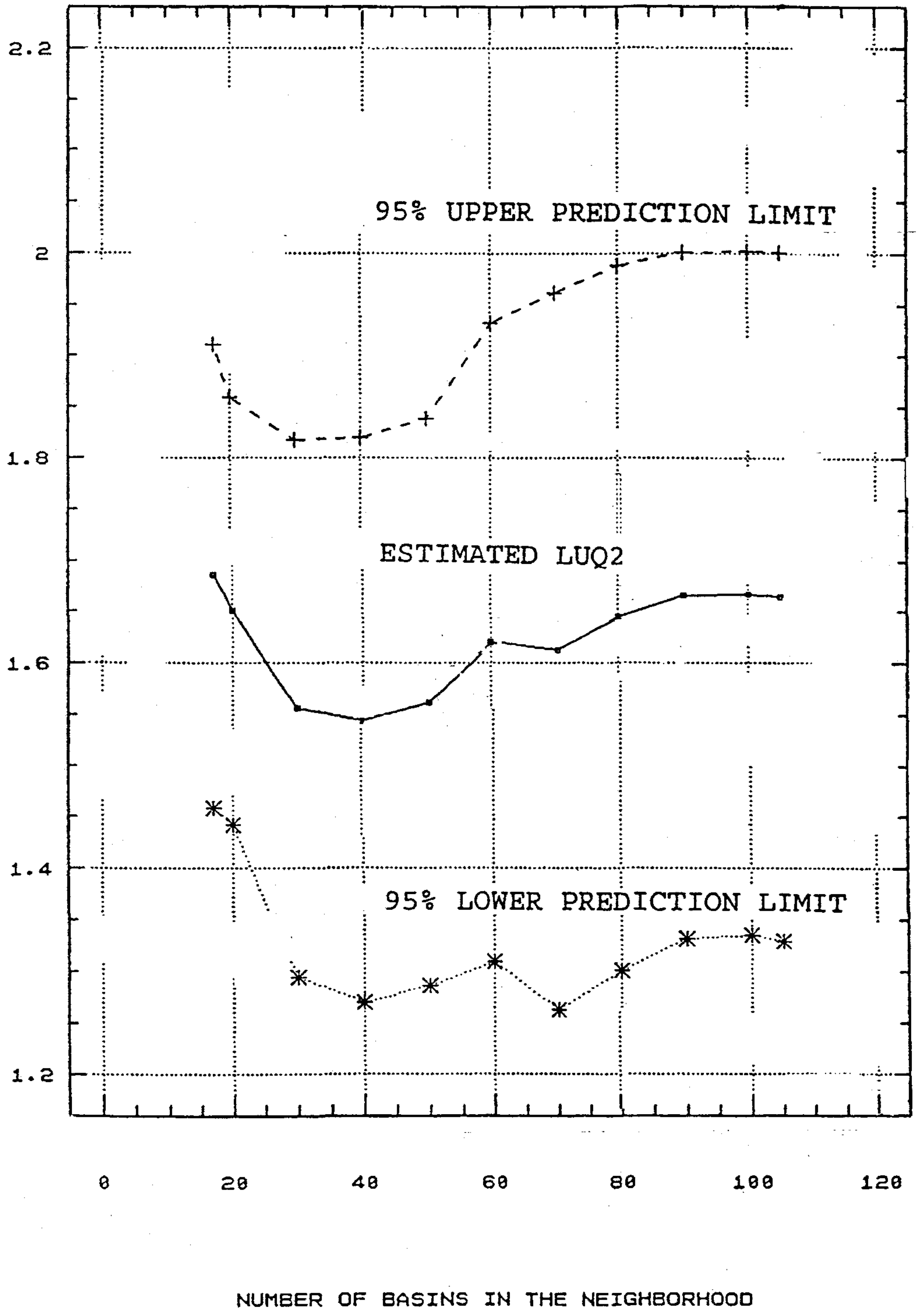
2.2
2
1.8
1.6
1.4
1.2
95% UPPER PREDICTION LIMIT
ESTIMATED LUQ2
95% LOWER PREDICTION LIMIT
0
20
40
60
80
100
120
NUMBER OF BASINS IN THE NEIGHBORHOOD

FIG. 9

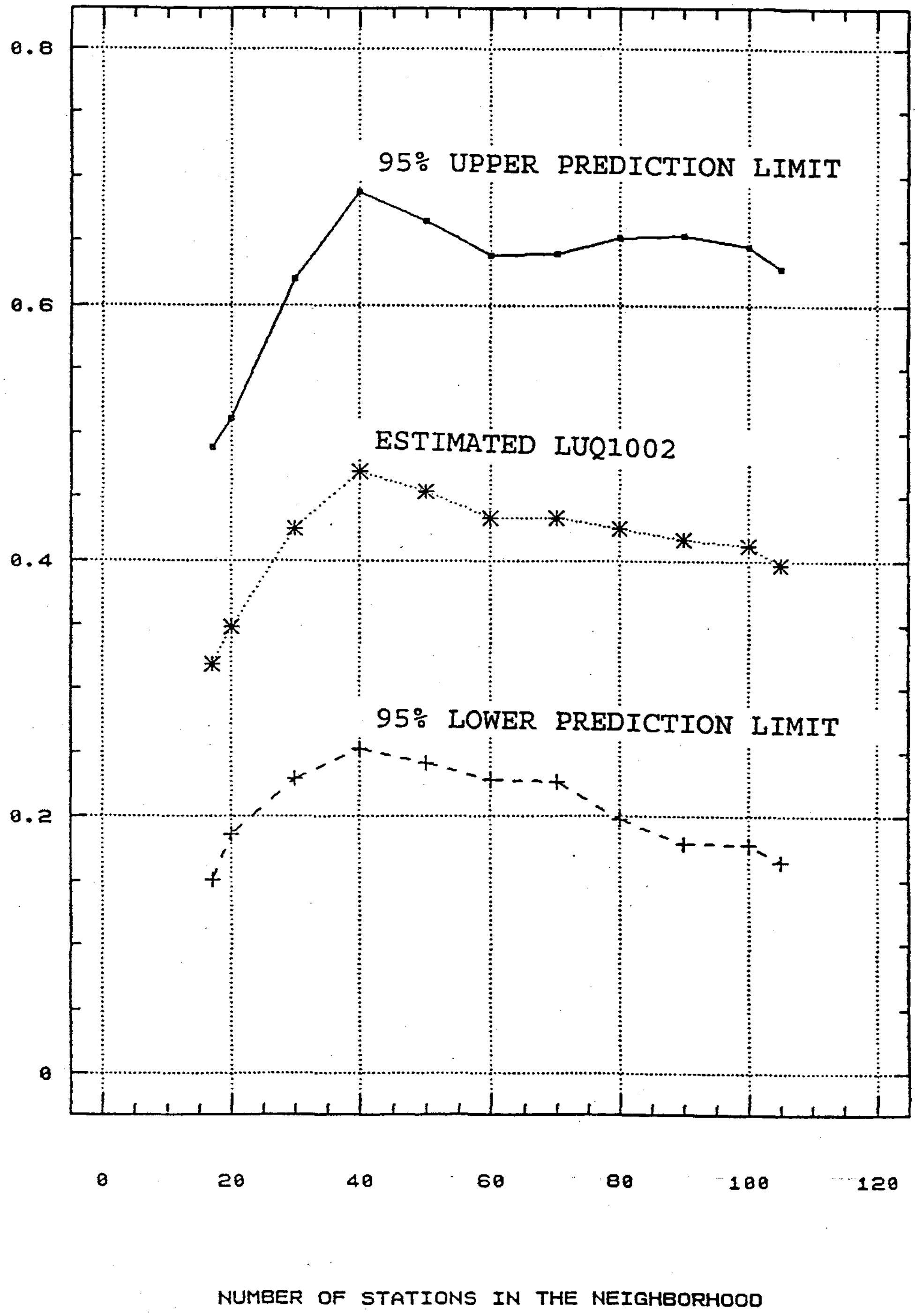
0.8
0.6
0.4
0.2
0
95% UPPER PREDICTION LIMIT
ESTIMATED LUQ1002
95% LOWER PREDICTION LIMIT
0
20
40
60
80
100
120
NUMBER OF STATIONS IN THE NEIGHBORHOOD

FIG. 10

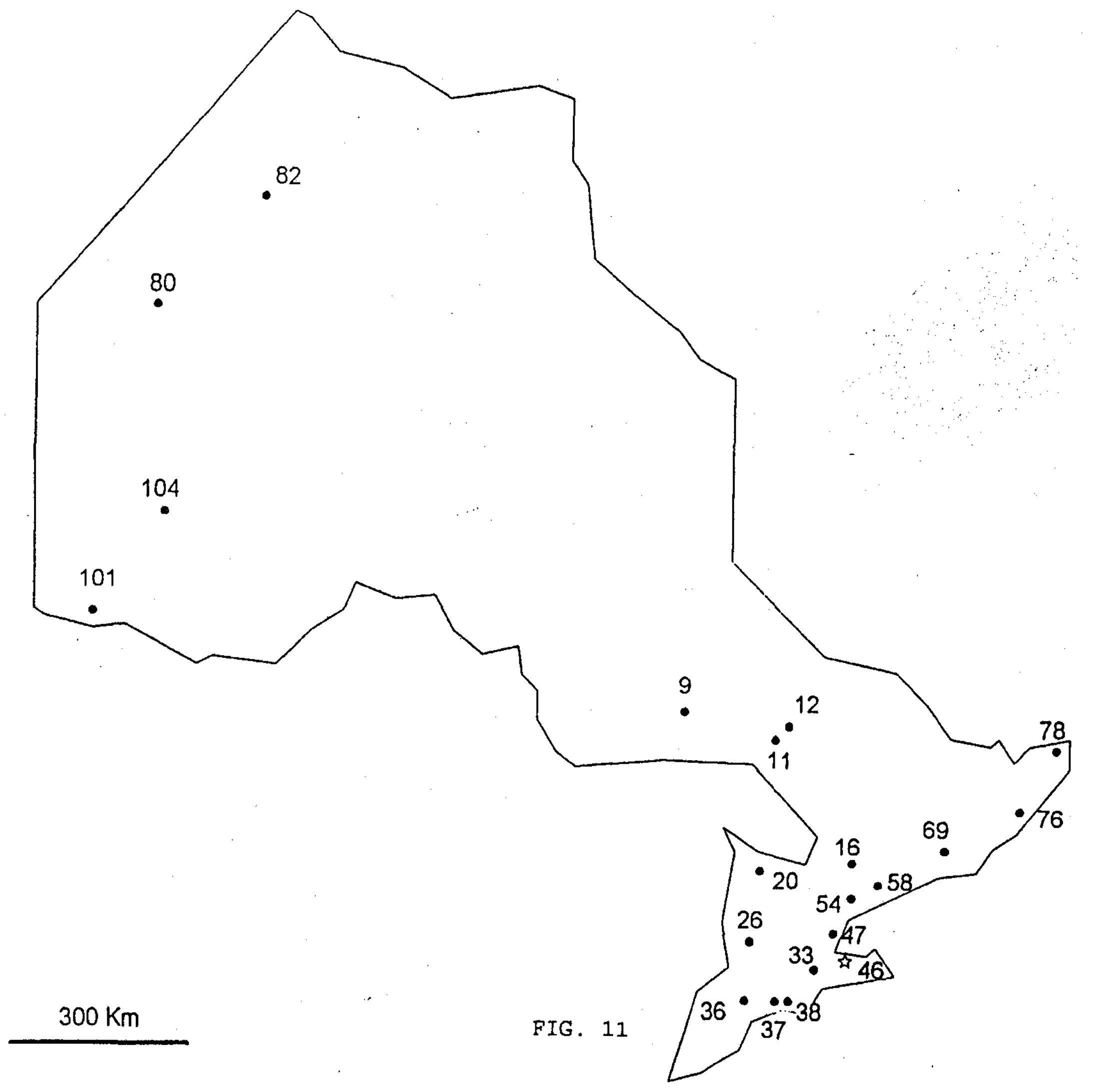

FIG. 11

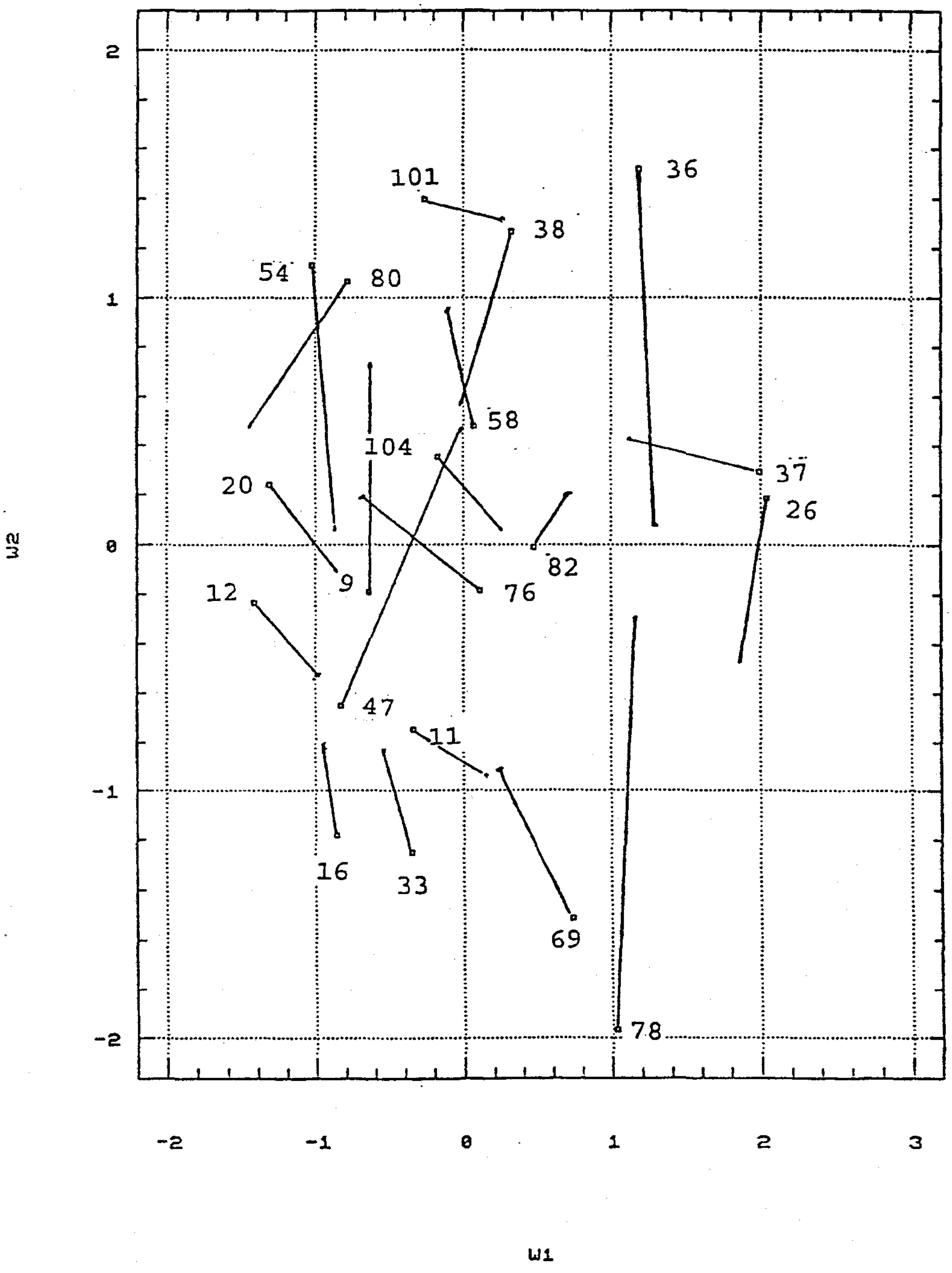
101
36
38
54
80
58
104
37
20
26
82
9
76
12
47
11
16
33
69
78
2
1
0
-1
-2
-2
-1
0
1
2
3
W2
W1

FIG. 12

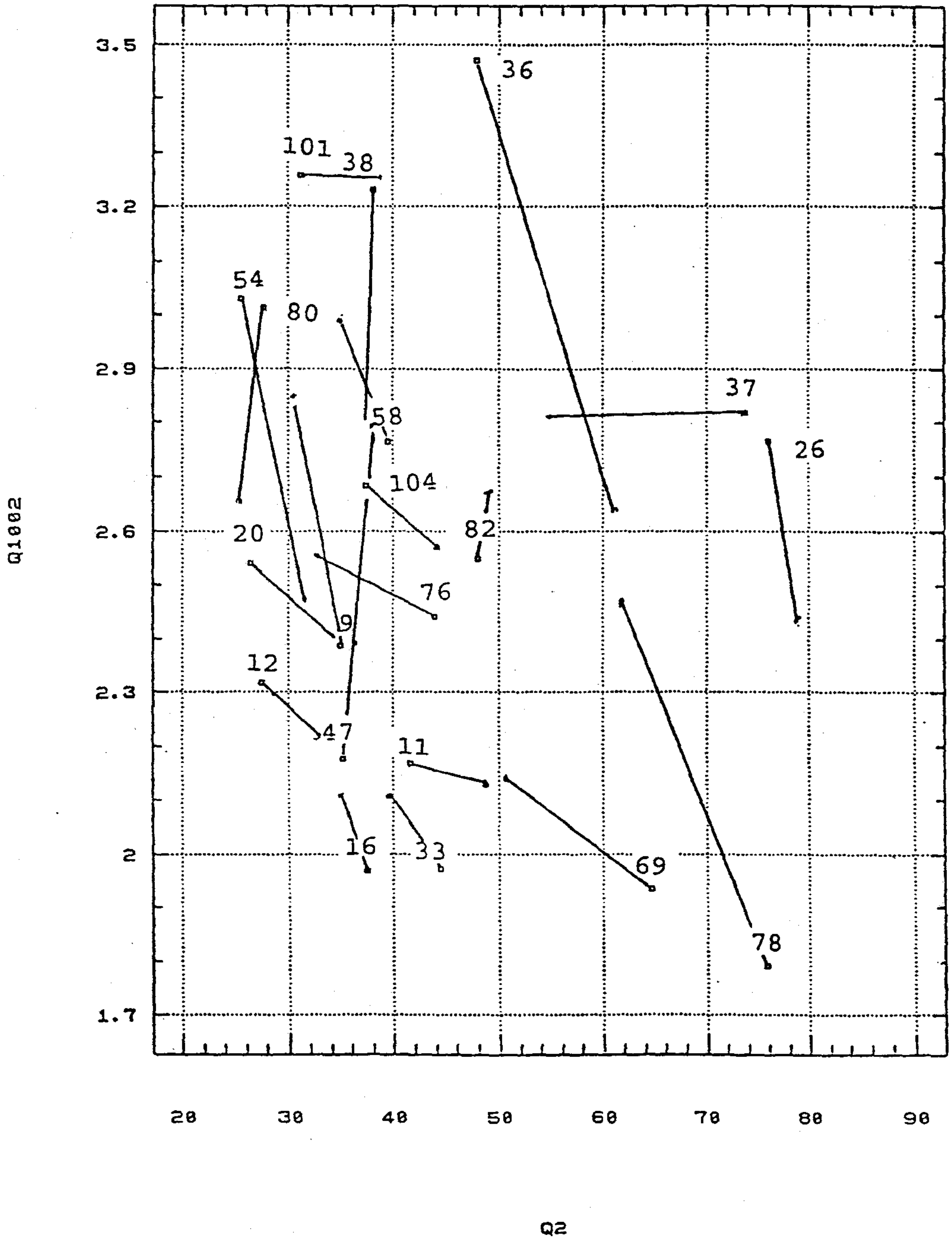
3.5
3.2
2.9
2.6
2.3
2
1.7
Q1002
20
30
40
50
60
70
80
90
Q2
36
101
38
54
80
58
104
82
20
76
9
37
26
12
47
11
16
33
69
78

FIG. 13