Hydrological Sciences Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/thsj20

Multivariate analysis in hydrology: the factor correspondence analysis method applied to annual rainfall data

L. SILVEIRA [a]

[a] Department of Fluid Mechanics and Environmental Engineering, School of Engineering, University of Uruguay, Julio Herreray Reissig 565, CP 11300, Montevideo, Uruguay

Available online: 25 Dec 2009

To cite this article: L. SILVEIRA (1997): Multivariate analysis in hydrology: the factor correspondence analysis method applied to annual rainfall data, Hydrological Sciences Journal, 42:2, 215-224

To link to this article: http://dx.doi.org/10.1080/02626669709492021

Multivariate analysis in hydrology: the factor correspondence analysis method applied to annual rainfall data

L. SILVEIRA
Department of Fluid Mechanics and Environmental Engineering, School of Engineering, University of Uruguay, Julio Herrera y Reissig 565, CP 11300 Montevideo, Uruguay

Abstract Most research on hydrology in developed countries does not emphasize problems that are often encountered by hydrologists in developing countries, problems such as the scarcity and poor quality of data. In developing countries the available data base is often a major constraint and a limiting factor in any hydrological study. This paper analyses the consistency of a network of nonrecording raingauges based on the annual rainfall recorded by six representative stations in the Tacuarembó River basin in Uruguay. The identification of possible errors not revealed in the printed records was accomplished by application of the factor correspondence analysis method and verified with success by the classical double-mass analysis. This ability to identify suspicious data with very modest requirements for data management implies that the methodology based on factor correspondence analysis could find wide application.

Analyse multivariée en hydrologie: l'analyse de correspondance appliquée à des données de précipitations annuelles

Résumé La plupart des travaux de recherche en hydrologie menés dans les pays développés ne s'intéressent pas beaucoup aux problèmes que rencontrent habituellement les hydrologues des pays en voie de développement, comme la rareté ou la médiocre qualité des données. Dans ce pays les données de base disponibles sont souvent un facteur de contrainte et de limitation majeur dans toute étude hydrologique. Cet article analyse la cohérence d'un réseau de pluviomètres: nous avons considéré les précipitations annuelles enregistrées en six stations représentatives du bassin de la rivière Tacuarembó en Uruguay. L'identification d'erreurs possibles non découvertes dans les enregistrements imprimés a été réalisée par l'analyse de correspondance, puis vérifiée avec succès en appliquant la méthode classique de l'analyse de doubles cumuls. Cette capacité à identifier des données douteuses, avec des exigences très modestes en ce qui concerne le traitement des donnés, signifie que la méthodologie pourrait trouver un vaste champ d'application.

INTRODUCTION

The increased availability of high speed computers has led to a proliferation of improved hydrological model structures that simplify reality by using a variety of mathematical approaches (Anderson & Burt, 1985; Moore, 1986). However, hydrological research in developed countries does not emphasize problems such as the scarcity and poor quality of input data which often affect hydrologists in developing countries, (Mutreja, 1986). The prevailing research scenarios in developing countries are based on data collecting networks and transmission equipment which are much more elementary than in developed countries. Noise errors due to transmission techniques as well as measurement errors due to the

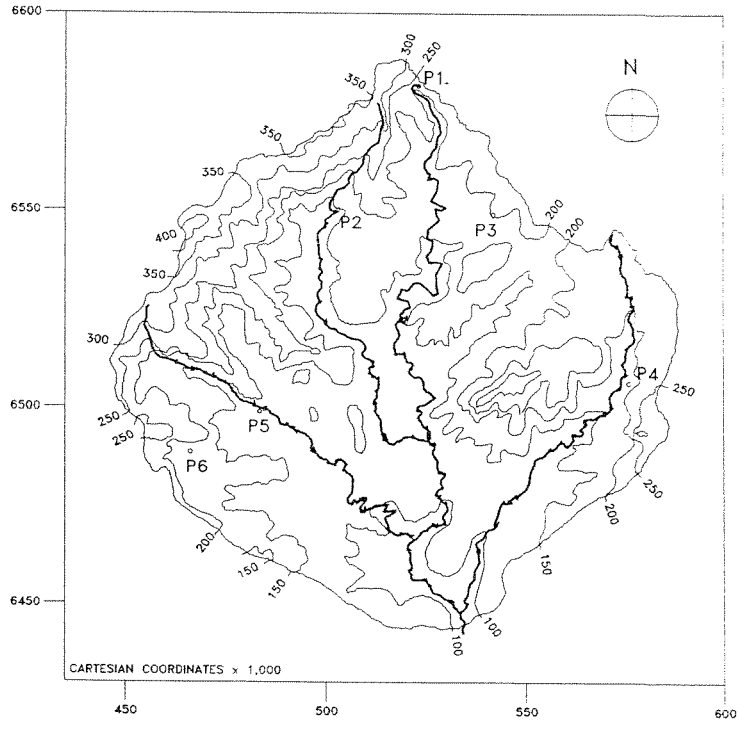

Fig. 1 Nonrecording raingauges in the Tacuarembó River basin in Uruguay.

instrumentation or the observer and typing errors due to the model operator are more common. Therefore, the quality and the length of the historical accessible data base are often a major constraint and a limiting factor in any hydrological study in developing countries (Tucci, 1986).

The success of any hydrological forecasting depends upon the quality of the input data. Calibration and validation of a hydrological model must be preceded by a quality analysis of input data. This paper shows an application of factor correspondence analysis to the study of the reliability of a network of nonrecording raingauges in the Tacuarembó River basin in Uruguay (Fig. 1).

STUDY BASIN

The Tacuarembó River basin in Uruguay has a surface area of 13 945 km^2 measured at the streamgauge section in Paso de la Laguna I. The time of concentration computed according to the Kirpich formula (Chow *et al.*, 1988) is 5.8 days. Along the northwest sector, the water divide consists of the Cuchilla Negra and the Cuchilla de Haedo, a mountain ridge with an altitude of about 250-400 m a.m.s.l. However, the landscape in the remaining area of the basin is slightly undulated with an altitude of about 100-200 m a.m.s.l. The main geological formations of the basin are essentially sedimentary deposits. A sand formation outcropping on the south is

Table 1 Nonrecording raingauges from the basic national network in the Tacuarembó River basin in Uruguay.

Station	Place	Reader	*X*	*Y*
P1 1147	Rivera	Meteorology	524.0	6582.5
P2 1220	Tranqueras	Home Office	503.0	6550.5
P3 1224	Ataques	Home Office	542.0	6549.0
0P4 1379	Moirones	Home Office	577.6	6504.0
P5 1405	Tacuarembó	Meteorology	482.5	6492.0
P6 1440	Valle Eden	Home Office	464.5	6480.0

identified as a recharge domain of a confined aquifer extending to Brazil, Argentina and Paraguay. Land cover is mainly natural pasture assigned to livestock. Rice farming has been increasing along the Tacuarembó River in recent years.

BASIC DATA

Six representative nonrecording raingauges inside the basin were chosen. These stations belong to the basic national network, which comprises 100 raingauges with records for the 65-year period from 1914 to 1978. The stations used are listed in Table 1 and shown in Fig. 1.

THE FACTOR CORRESPONDENCE ANALYSIS METHOD

In a review of the literature on the factor correspondence analysis method, Jackson (1991) points out that much of the early development as well as the application of this method were carried out in France, motivated primarily by the work of Benzecri (1973, 1977). The first text devoted primarily to this subject was by Lebart *et al.* (1977). This text contains many FORTRAN programs with which to carry out factor correspondence analysis. Another text on the subject is by Greenacre (1984) and Lebart *et al.* (1984).

Factor correspondence analysis differs from other multivariate methods such as principal component analysis in that instead of subtracting out row and/or column means, the matrix is multiplied by functions of the row and column totals, obtaining what is sometimes referred to as a chi-square metric (Jackson, 1991). There are several methods proposed to do this but among the most widely used is the procedure described in Greenacre (1984).

Factor correspondence analysis produces a biplot by taking into account the proportion of the total variability explained by the first two factors. Biplots, as originally conceived by Gabriel (1971, 1981), meant two-dimensional plots representing the first two factors summarizing the information contained in both variables and observations. Biplots are easy to work with because their interpretability surpasses the loss of original information. However, the two-dimensional synthesis of points that originally are contained in a multidimensional

space may be misleading. For example, two observations or variables close to each other in a two-dimensional biplot may be situated at a great distance in a n-dimensional space (with regard to the n observations) or a p-dimensional space (with regard to the p variables). The work of Benzecri (1973, 1977) overcame this problem by a proposition about the contributions which are coefficients that help to decide how reliably an observation or a variable has been restored in a limited factorial space. These coefficients are both the absolute contribution and the relative contribution. The absolute contribution represents the contribution of variables/observations to each factor. The relative contribution shows the contribution of each factor to the separation of variables/observations with respect to the average location (the centre of gravity of both clouds). The restitution of variables/observations in the plane of the first two factors is satisfactory when the sum of the relative contributions of each variable/observation is close to 100%.

CORRESPONDENCE ANALYSIS APPLIED TO ANNUAL RAINFALL DATA

Given a data matrix of n (rows) $\times$ p (columns), where n is the number of annual rainfalls considered as observations and p is the number of raingauges considered as variables, the resulting biplot obtained by applying factor correspondence analysis may be interpreted as follows:

Observations (annual rainfall)

- Each point representing an observation (year) contains the whole information (annual rainfall) recorded in p raingauges. This means that, p records are condensed to a single point.
- The distance between points representing different observations (years) is a measure of similitude with regard to annual rainfall on a given basin or region.

Variables (raingauges)

- Each point representing a variable (raingauge) contains the entire information (annual rainfall) recorded during a total of n observations (years). This means that, n records are condensed in a single point.

  If the points representing raingauges are situated close to each other, the relevant raingauges show a similar behaviour along the n observations (years). Otherwise, if the points are situated away from each other, the raingauges show a different behaviour.

  If the data matrix is free from errors, the factor correspondence analysis allows one to establish a regionalization. On the other hand, if it is not possible to admit that the data matrix is free from errors, a point situated away from the others may represent a raingauge that may be suspected of containing errors.

Relationship between observations (annual rainfall) and variables (raingauges)

A relationship between the two data groups (raingauges and annual observations) may also be attempted. In general, each point representing an observation will be situated in a position near a raingauge where the precipitation was large while those points situated on the opposite side indicate that in the considered year the raingauges registered little precipitation.

The above interpretation of the resulting biplot in a limited factorial space is valid, however, only if a reliable restitution of the n-dimensional and p-dimensional space is achieved. The accuracy of this restitution may be analysed by controlling the contributions according to the work of Benzecri (1973, 1977).

RESULTS AND DISCUSSION

The annual rainfall data recorded by six representative nonrecording raingauges inside the Tacuarembó River basin were analysed. The factor correspondence analysis was computed using a FORTRAN program following the procedure described in Greenacre (1984). The variance explained by the first two factors was 68.12%. The sum of the relative contributions of raingauges/annual rainfall was close to 100%. Therefore, the restitution in a two dimensional space is satisfactory according to Benzecri (1973, 1977). The resulting biplot (Fig. 2) shows that raingauges P3 (Artigas), P4 (Moirones) and P6 (Valle Eden) are situated near the centre of gravity (0, 0). Thus, according to the interpretation of the factor

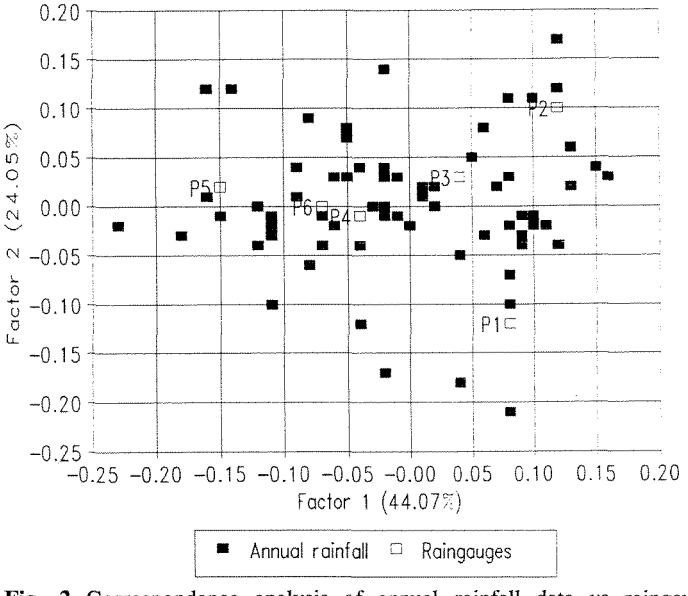

Fig. 2 Correspondence analysis of annual rainfall data *vs* raingauges for the Tacuarembó River basin.

correspondence analysis described above, these stations represent the average behaviour in the study basin.

The remaining raingauges, P1 (Rivera), P2 (Tranqueras) and P5 (Tacuarembó) are situated far from the centre of gravity (0, 0). Raingauge P1 (Rivera) is located on the far north edge of the basin, close to the Cuchilla Negra, a mountain ridge of about 300 m a.m.s.l. This means that observed higher rainfall may be explained by topographical influences.

P2 (Tranqueras) is located somewhat further from the centre of gravity (0, 0) than is P3 (Artigas). Both stations are situated at about the same latitude. However, the topographical map shows that P2 is located near the Cuchilla de Haedo, another mountain ridge of about 300-400 m a.m.s.l. altitude

The distance between P5 (Tacuarembó) and P6 (Valle Eden) is 21.6 km. Thus, these raingauge stations are relatively close to each other. Moreover, there is no significant topographical difference between their locations. Thus the question is how to explain their different behaviour according to correspondence analysis. As a first step to account for this difference, the original data were analysed. Annual data from P6 was simply subtracted from P5 data to produce Fig. 3. An inspection of Fig. 3 shows that between 1931 and 1956 P5 data were always greater than those of P6. There were differences of about 700 mm during two consecutive years, which can be compared with the average annual rainfall in the Tacuarembó River basin of about 1300 mm. In 1947 the annual rainfall at P5 was 63% greater than the annual rainfall at P6 (1418 mm against 869 mm). The question arises as to whether such systematic differences between two raingauges situated close to each other could be possible. Assuming that possible errors in records may be related to the station with the higher precipitation, the next step in the investigation was to carry out a double-mass

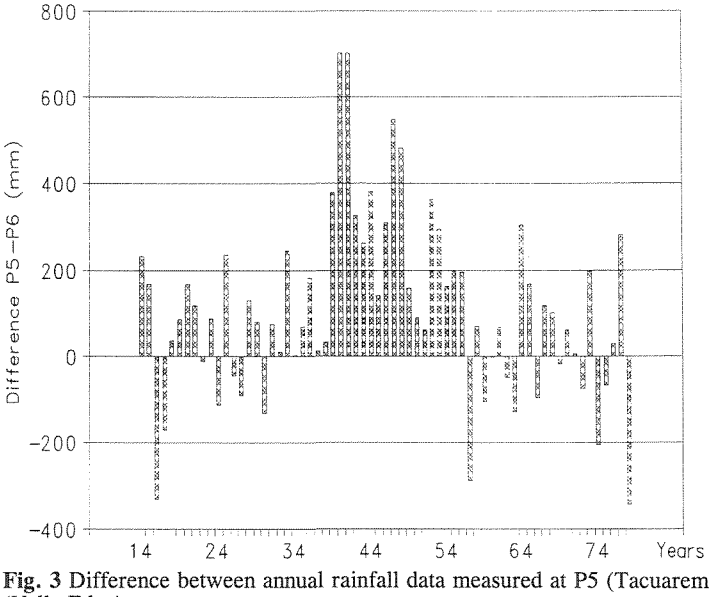

Fig. 3 Difference between annual rainfall data measured at P5 (Tacuarembó) and P6 (Valle Eden).

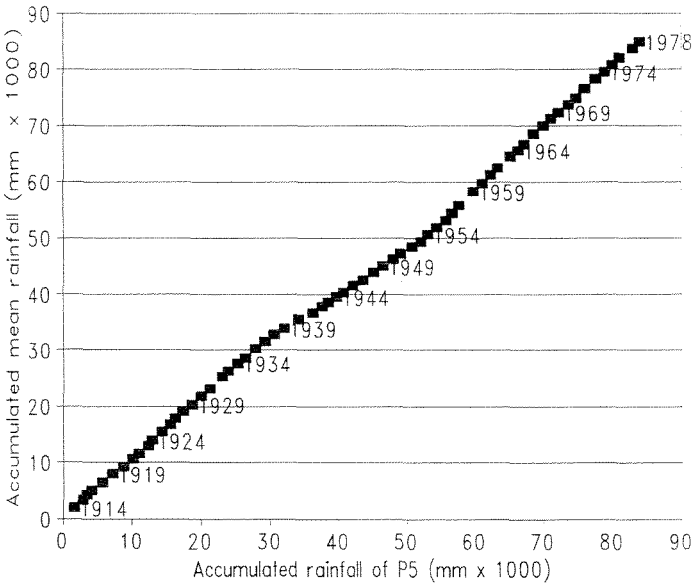

Fig. 4 Adjustment of rainfall data for P5 (Tacuarembó) by double-mass curve.

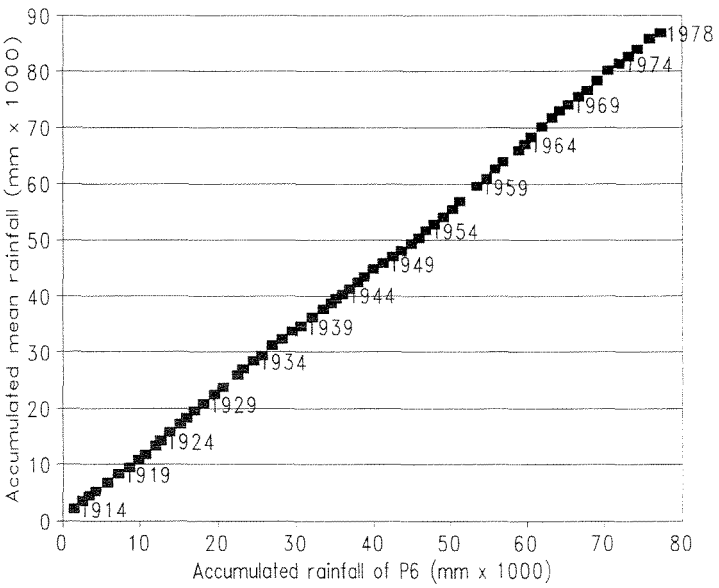

Fig. 5 Adjustment of rainfall data for P6 (Valle Eden) by double-mass curve.

Adjustment of rainfall data for P5 (Tacuarembó) by double-mass curve. analysis (Chow, 1964) to test the consistency of the records measured at P5. Double-mass analysis was accomplished comparing the accumulated annual rainfall of P5 with the corresponding accumulated values of average rainfall measured in the other group of raingauges (P1, P2, P3, P4 and P6). Two changes in the slope are apparent in the resulting plot (Fig. 4). The first change was in 1937 and the second in 1952. The

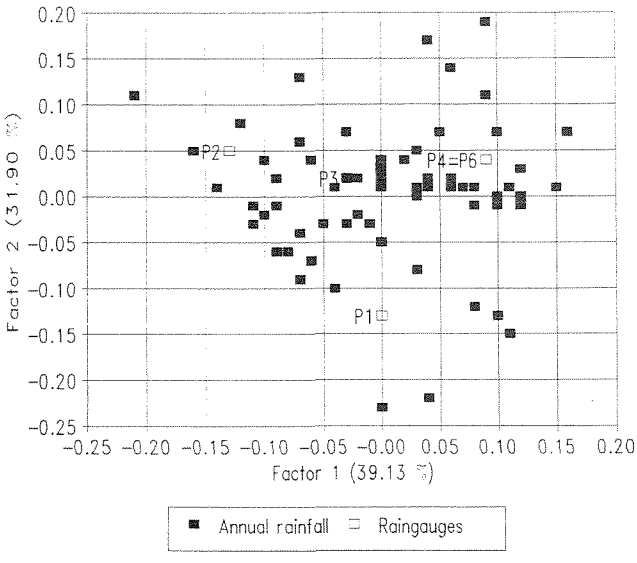

Fig. 6 Correspondence analysis of annual rainfall *vs* remaining raingauges by eliminating P5 (Tacuarembó).

mean lines before 1937 and after 1952 are parallel. According to Heras (1976), this type of change in a slope may be interpreted as a systematic measurement error that has been made during the central period between 1937 and 1952. To check this statement, the consistency of the records of the base stations was tested. A double-mass analysis between P6 and the remaining stations (P1, P2, P3 and P4) showed that the plotted points fall on a mean line (Fig. 5).

Raingauge P5 was subsequently eliminated and factor correspondence analysis was carried out again for P1, P2, P3, P4 and P6. The variance explained by the resulting biplot (Fig. 6) increased from 68.12% to 71.03%. Raingauge P3 is located close to the average point (0, 0) while P1, P2 and P4, P6 are located at about the same distances from the average point but in different directions. This suggests a spatial variability of rainfall in the Tacuarembó River basin. Examination of all the annual data shows that annual rainfall decreases from northwest to southeast. The mean annual rainfalls at P1, P2, P3, P4 and P6 for the 65 years period considered are 1453, 1327, 1326, 1235 and 1180 mm respectively. Decreasing latitude and decreasing altitude are associated with decreasing annual rainfall.

The example above shows that two suspect raingauges (P5 and P6) were identified directly in the resulting biplot. However, the example also shows that two other raingauge (P1 and P2) were discarded as suspect due to differences in topographical altitude. Differences in physiography and climate must be taken into account by the hydrologist before a raingauge is classified as suspect due to its location in a biplot.

CONCLUSIONS

The application of any multivariate method presumes the basic hypothesis that the data matrix is free from errors. A literature review about multivariate methods (for

example, empirical orthogonal functions, principal component analysis and factor analysis) showed that more than a hundred applications were published in the meteorological literature during the last decade. Eigenvalue techniques have been used to address such diverse objectives as, for example, pure data reduction and climate regionalization studies regarding temporal and spatial variability of rainfall. This paper has shown that the factor correspondence analysis method can be used to study the consistency of annual rainfall data recorded by a network of nonrecording raingauges and to identify possible errors not revealed in the printed records. Whether other multivariate methods could be adopted as techniques to detect errors is a matter for further research. The ability of factor correspondence analysis to detect outliers has been demonstrated by analysing the annual rainfall data recorded by six representative nonrecording raingauges in the Tacuarembó River basin. Two raingauges situated close to each other were identified as likely to contain errors. An analysis of the original data showed a systematic difference during a period of 16 years. Subsequently, the existence of systematic measurement errors was verified with success by applying classical double-mass analysis. The application of factor correspondence analysis also showed that this method is faster than the classical double-mass method, which requires a separate analysis for each raingauge in a network. Nevertheless, double-mass analysis is necessary to decide which of any suspect raingauges are erroneous. However, the computations of double-mass analysis are reduced only to the suspect raingauges identified by just one computation with correspondence analysis. The study presented in this paper dealt with only six raingauges. Correspondence analysis can be advantageous when the number of raingauges is larger.

The proposed methodology based on the factor correspondence analysis method could be easily applied to study the consistency of a network of nonrecording raingauges in other river basins.

Acknowledgements The author gratefully acknowledges Dr E. Usunoff who introduced him to the theory of factor correspondence analysis and its applications in hydrology.

REFERENCES

Anderson, M. G. & Burt, T. P. (1985) Modelling strategies. In: Hydrological Forecasting (ed. by M. G. Anderson, & T. P. Burt). Wiley, Chichester, UK.

Benzecri, J. P. (1973) *L'Analyse des Données*, II: *L'Analyse des Correspondances* .Dunod, Paris, France (in French).

Benzecri, J. P. (1977) Histoire et prehistoire de l'analyse des données, V. L'analyse des correspondences. *Cahiers de l'analyse des données* (in French).

Chow, V. T. (1964) *Handbook of Applied Hydrology* McGraw-Hill, New York, USA.

Chow, V. T., Maidment, D. R. & Mays, L. W. (1988) *Applied Hydrology*. McGraw-Hill International Editions, New York, USA.

Gabriel, K .R. (1971) The biplot-graphic display of matrices with application to principal component analysis, *Biometrika* **58**, 453-467.

Gabriel, K. R. (1981) Biplot display of multivariate matrices for inspection of data and diagnosis. In: *Interpreting Multivariate Data* (ed. by V. Barnett), 147-173. John Wiley, New York, USA.

Greenacre, M. J. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press, London, UK.

Heras, R. (1976) Hidrologia y Recursos Hidráulicos. Servicio de Publicaciones del Ministerio de Obras Públicas, Madrid, Espaná (in Spanish).

Jackson, J. E. (1991) *A User's Guide to Principal Components*. Wiley, New York, USA.

Lebart, L., Morineau, A. & Tabard, N. (1977) *Techniques de la Description Statistique: Mètodes et Logiciels Pour L'analyse des Grands Tableaux*. (in French) Dunod, Paris, France.

Lebart, L., Morineau, A. & Warwick, K. (1984) *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. Wiley New York, USA.

Moore, R. J. (1986) Advances in real-time flood forecasting practice. In: *Symp. on Flood Warning Systems*, (Winter Meeting of the River Engineering Section, The Institution of Water Engineers and Scientists).

Mutreja, K. N. (1986) *Applied Hydrology*. Tata McGraw-Hill, New Delhi, India.

Tucci, C. E. (1986) Modelos Matemàticos em Hidrologia e Hidráulica. *Revista Brasileira de Engenharia* (RBE) (3 volumes).

Received 13 March 1996; accepted 21 October 1996