SPECTRAL DECOMPOSITIONS IN CLUSTER ANALYSIS WITH APPLICATIONS TO LIMNOLOGICAL DATA

LUIS MAURICIO BINI *; JOSÉ ALEXANDRE FELIZOLA DINIZ-FILHO **

*Escola de Engenharia de São Carlos EESC/USP
Centro de Recursos Hídricos e Ecologia Aplicada - CRHEA
Av. Dr. Carlos Botelho, n. 1465
Caixa Postal 359, CEP: 13.560 – São Carlos – SP – Brasil
**Departamento de Biologia Geral, ICB
Universidade Federal de Goiás - Campus II
Caixa Postal 131, CEP: 74.001-970 – Goiânia – GO – Brasil

RESUMO: Decomposição espectral de análise de agrupamento aplicada a dados limnológicos. A formação de grupos baseados em dendrogramas obtidos por análise de agrupamento é feita usualmente de maneira arbitrária. Neste trabalho é apresentado um método para detectar o nível de formação dos grupos onde há uma maximização da correlação entre a matriz de distâncias originais e uma matriz modelo derivada do dendrograma. O método é ilustrado com um dendrograma obtido a partir de variáveis limnológicas de rios do litoral Sul Paulista. A classificação final obtida é congruente com as espectativas baseadas no conhecimento ecológico do sistema e em seu nível de poluição.
PALAVRAS-CHAVE: Análise de agrupamento, classificação, teste de Mentel, nível de corte.

ABSTRACT: Spectral decompositions in cluster analysis with applications to limnological data. The establishment of groups based on dendrograms obtained through hierarchical cluster analysis is usually done in an arbitrary way. In this paper, it is presented a method to detect at which level of clustering there is a maximization of the correlation between the calculated distance matrix among sampling unities (or operational taxonomic unities – OTU's) and a model matrix obtained from the dendrogram. The method is ilustrated by grouping sampling stations of six streams in the state of São Paulo (Brazil) by ten physical-chemical parameters of the water. The final classification obtained is correspondent to the expectations based on the ecological status of the system and its level of pollution.
KEY WORDS: Cluster analysis, classification, Mentel test, cut level.

INTRODUCTION

Cluster analysis is an algebric method which allows the establishment of groups from a distance or similarity matrix among pairs of objects (Sneath & Sokal, 1973). Sequential, agglomerative, hierarchical and non-overlapping clustering methods (SAHN) have been extensively used in systematic (Sneath & Sokal, 1973; Buth, 1984) and ecology (Legendre & Legendre, 1983; Krebs, 1989). In Brazil, several recent limnological studies (Tolentino *et al.*, 1986; Huszar & Esteves, 1988; Lanzer & Schäfer, 1988; Lobo *et al.*, 1992; Huszar & Silva, 1992) used these techniques to assess the relationships among sampling stations based on biotic and/or abiotic data.

The outcome of cluster hierarquical analysis is displayed by tree like diagrams, called dendrograms. These dendrograms can be partitioned in levels of distance or similarity, being necessary to choose one of them as cut off point to establish the number groups of objects. This choice, however, has been made in an arbitrary way, sometimes based on theoretical conceptions established *a priori*.

In this paper it is presented a method for the decomposition of a dendrogram through the analysis of its different sections. This analysis is performed by comparing by how much the distances among groups of objects are larger than the distances within these groups. This decomposition, therefore, should indicate at which level of the dendrogram there is a maximization of the original distances among groups, which is suitable to establish biological or ecological classifications.

METHODS

Depending on the nature of the original data, dendrograms are constructed based on a distance or similarity matrix between pairs of n objects (Sneath & Sokal, 1973; Krebs, 1989; Gower & Legendre, 1986).

The coefficient of cophenetic correlation is obtained by comparing the original matrix (n x n) of distances or similarity with a resemblance matrix between pair of objects obtained from the dendrograms. It allows the evaluation of how much the clustering processes change the patterns of multidimensional resemblance between the objects. The computation of this coefficient is usually the only evaluation performed in dendrograms, and it does not allow the definition of the number of groups of objects that can be established.

The method presented in this paper is based on the computation of the cophenetic correlation coefficient and in model matrices derived from dendrograms (Rohlf, 1974; Manly, 1986; 1991). The main idea is to obtain several correlation coefficients between matrices by comparing the original patterns of distances or similarities between the objects with several model matrices (Manly, 1991), as follows: the dendrogram is partitioned in several levels, and for each one a model matrix is constructed confering the value of 1.0 if the pair of objects being compared is in the same group and the value of zero otherwise.

The correlations between matrices obtained measure, therefore, the fit of the original distance or similarity matrix between the objects and the models, defining several patterns of resemblance. The significance of each correlation can be tested using the Mantel test (Manly, 1991). The Mantel Z statistic is given by:

$$Z = \Sigma_i \Sigma_j (X_{ij} . Y_{ij})$$

where X_{ij} and Y_{ij} are, respectively, the elements of the matrices **X** and **Y** being compared. The null hypothesis is the independence between matrices (correlation is zero). The significance of the Z value against the null hypothesis is established, then, by recomputing it a given number of times by randomly permuting the order of the rows and the columns of one of the matrices (Manly, 1991). In the present study the model matrix is the independent one (matrix **X**).

The relationship between the correlation coefficients obtained and the levels of section of the dendrogram can be considered similar to a spatial correlogram (Sokal & Oden, 1978), or as a spectre of the relationships among the elements of the dendrogram. Because of this the proposed method has been called spectral decomposition of cluster analysis.

APPLICATION OF THE PROPOSED METHOD

To illustrate the proposed method, it was used data on chemical composition of water of 13 sampling stations (objects) of 6 rivers of São Paulo State (Brazil) located in the Itanhaém and Cananéia Basin. The variables analysed were: temperature, pH, alkalinity, salinity, CO_2, O_2, total-N, total-P, and organic Carbon . More details about data gathering and it ecological interpretation can be found in Bini (1991).

The measure between pairs of sampling stations used was the average euclidian distance, computed after the standardization of the data. The dendrogram was obtained using UPGMA (unweighted pair group method - arithmetic average) (Sneath & Sokal, 1973).

The dendrogram obtained was arbitrarily sectioned in 13 levels, at intervals of distance equal to 0.25, allowing the construction of 13 model matrices. The model matrices 1 and 13 have, respectivelly, all values equal to 1.0 and 0.0, since at level 1 all the sampling stations form a single group, and at the level 13 each sampling station forms a group. Those two matrices were excluded from the analysis. The correlations between the matrices were tested with the Mantel test. One thousand random permutations of each matrix were run, using the routine MXCOMP from NTSYS-PC, version 1.5 (Rohlf, 1989).

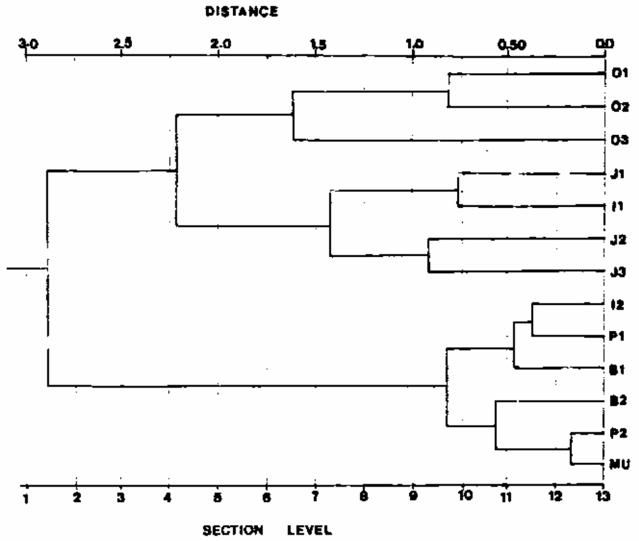

Figure 1 - Dendrogram of classification of 13 sampling stations and section levels established for this study.

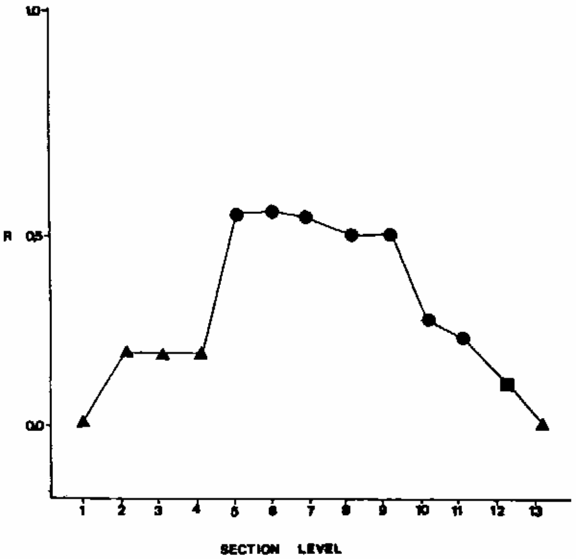

Figure 2 – Correlation between matrices (R) obtained from different section levels of the dendrogram (ns ▲; $P < 0.05$ ■; $P < 0.01$●).

RESULTS

The dendrogram obtained using UPGMA, already sectioned, is given in the fig. 1. The cophenetic correlation coefficient was 0.806. The correlations between matrices obtained and their significance levels computed show that between levels 5 and 6 (distances equal to 1.75 and 2.00) is the highest correspondence between the original distances and the groups established by the dendrogram (fig. 2). At those levels, thus, there is a maximization of the average euclidean distances among groups in relation to the within groups distances. As a result, 3 groups of sampling stations can be observed: A) O1, O2 and O3; B) J1, I1, J2 and J3 and C) I2, P1, B1, B2, P2 and Mu. The fig. 2 shows the relation between the matrix correlations obtained and the levels of section. However, it is important to note that significant correlations between matrices ($P < 0.05$) were obtained for sections 5 to 12.

DISCUSSION AND CONCLUSIONS

The groups formed at levels 5 and 6 are the most representative regarding the maximization of distance among the groups of sampling stations. The groups defined at these levels correspond to the classification of rivers in endogenous (J1, J2, I1, J3) and exogenous (I2, P1, B1, P2, B2, Mu) proposed by Navarra (1988). The sampling stations of the Olaria river (O1, O2, O3), considered endogenous, formed an isolated group because of the strong discharges of organic sewer in this river (Bini, 1991). In this way, the ecological interpretation of the data demonstrates the relevance of the analytic method proposed.

At levels 5 and 6 occur the highest correspondance between the original distances and the model. However, other levels, such as level 8, can also give information about the pattern of resemblance. At this level, the correlation between matrices was 0.564 ($P<0.001$) and allows the analysis of a heterogenity among the sampling stations of the Olaria stream (O1 and O2 in a subgroup and O3 in another) and the group B, separating the sampling stations J1 and I1 from the sampling stations J2 and J3. This level, however, does not indicate a maximization level and must be considered in a secondary way. Other groups established at other levels, e. g. 2, 3 and 4 or 11 and 12, are not valid, since the distance among groups of stations are not significantly greater than the distances among the stations of another group.

In this example, the number of sections (13) of the dendrogram was established at intervals of average euclidean distance equal to 0.25, chosen arbitrarily, like the number of distances classes in spatial correlograms (Sokal & Oden, 1978). The number of levels of section to be established, however, must take into account the number of elements in the dendrogram and also the complexity of their linkages, allowing a more detailed definition of the ecological and biological relations among the elements considered.

ACKNOWLEDGEMENTS

We thank Prof. Dr. A.F.M. Camargo and two anonymous reviewers for their constructive criticisms on the manuscript.

REFERENCES

BINI, L. M. (1991). *Limnologia de alguns ecossistemas lóticos do litoral Sul Paulista: Aspectos físicos e químicos.* Rio Claro, UNESP. 68p. (Trabalho de Formatura).

BUTH, D. G. (1984). The application of electrofòretic data is systematic studies. *Ann. Rev. Ecol. Syst.*, 15: 501-522.

GOWER, J. C. & LEGENDRE, P. (1986). Metric and euclidean properties of dissimilarity coefficients. Journal of *Classification*, 3: 5-48.

HUSZAR, V. L. M. & ESTEVES, F. A. (1988). Considerações sobre o fitoplâncton de 14 lagoas costeiras do Estado do Rio de Janeiro, Brasil. *Acta Limnol. Brasil.*, II: 323-345.

HUSZAR, V. L. M. & SILVA, L. H. S. (1992). Comunidades fitoplanctônicas de quatro lagoas costeiras no norte do Estado do Rio de Janeiro, Brasil. *Acta Limnol. Brasil.*, VI: 291-314.

KREBS, J. C. (1989). *Ecological Methodology.* New York, Harper & Row, Publisher. 654p.

LANZER, R. M. & SCHÄFER, A. (1988). Fatores determinantes da distribuição de moluscos dulceaquícolas nas lagoas costeiras do Rio Grande do Sul. *Acta Limnol. Brasil.*, II: 649-675.

LEGENDRE, L. & LEGENDRE, P. (1983). *Numerical Ecology.* New York, Elsevier. 419p.

LOBO, E. A.; CALLEGARO, V. L. M.; FERRAZ, G. C.; ALVES-DA-SILVA, S. M. (1992). Análise da estrutura da biocenose de diatomáceas em lagoas da Estação Ecológica do Taim, Rio Grande do Sul, Brasil. *Acta Limnol. Brasil.*, VI: 277-290.

MANLY, B. J. F. (1986). *Multivariate Statistical Methods: A primer.* London, Chapman and Hall. 159 p.

_____ . (1991). Randomization and Monte Carlo Methods in Biology. London, Chapman and Hall. 281p.

NAVARRA, C. T. (1988). Fácies hidroquímicas dos rios da planície costeira Sul-Paulista. *Acta Limnol. Brasil.*, II: 931-942.

ROHLF, F. J. (1974). Methods of comparing classifications. Ann. Rev. Ecol. Syst., 5: 101-113.

_____ . (1989). NTSYS - *Numerical Taxonomy and Multivariate Analysis System.* New York, Exeter Publishing, LTD.

SNEATH, P. H. A. & SOKAL, R. R. (1973). *Numerical Taxonomy: The principles and practice of numerical classification*. San Francisco, W. H. Freeman and Co. 573 p.

SOKAL, R. R. & ODEN, N. L. (1978). Spatial autocorrelation in biology: 1. Methodology. *Biological Journal of the Linnean Society*, 10: 199-228.

TOLENTINO, M.; ESTEVES, F. A.; ROLAND, F.; THOMAZ, S. M. (1986). Composição química do sedimento lacustre de doze lagoas do litoral Fluminense e sua utilização na tipologia destes ecossistemas. *Acta Limnol. Brasil.*, I: 431-447.

Aceito em 12-05-94