A B S T R A C T A B O T A N I C A

operum ex Instituto Taxonomiae-Oikologiae Plantarum
Univ. Sci. de L. Eötvös nominatae

Tomus VI.

Redigit: T. SIMON

Budapest (Hungaria)

1980

Az Eötvös Loránd Tudományegyetem Természettudományi Kara
Növényrendszertani és Ökológiai Tanszékének Kiadványa
Tanszékvezető: Dr. Simon Tibor

SYN-TAX:
Számitógépes programcsomag ökológiai, cönológiai és taxonómiai osztályozások végrehajtására

Podani János
MTA Botanikai Kutató Intézete
Vácrátót

1980

SYN-TAX: Computer program package for cluster analysis in ecology, phytosociology and taxonomy

J. Podani

Research Institute for Botany of
the Hungarian Academy of Sciences
Vácrátót, Hungary 2163

SUMMARY

In the recent years a great deal of numerical classification techniques has been developed. These methods have enjoyed increasing popularity in the analysis of ecological, taxonomical and phytosociological data. One of the fundamental conditions of the wide-spread application of clustering methods is a computer program package available for all research workers. Although we can find numerous publications in the literature which contain listings and full documentations of classification programs, the author thought it necessary to develop an entirely original program system for two main reasons. First, three combinatorial algorithms and some techniques based on information theory have not been programmed so far. Secondly, Hungarian biologists have scarcely applied clustering methods to their data for lack of easily attainable program package.

From 1976 to 1978, during the course of a classification study of some rocky grassland communities, four computer programs were written is ANSI FORTRAN and run on a CDC 3300 computer. In 1979 all programs were rewritten to develop a more broadly applicable program system.In their present forms these programs may be applied not only to phytosociological data but also to any other type of multivariate data.

This system is called 'SYN-TAX'. All procedures programmed are hierarchical classification methods. Three programs include agglomerative techniques while the fourth one per-

forms the association analysis known as the most typical divisive method. The programs use 19 subroutines, seven out of which are called by two or three programs. Table 3.2.1. lists the subroutine requirement of each program.

The agglomerative programs (NCLAS, HMCL, INFCL) have the following characteristics in common:

(1) The general algorithm of the agglomerative methods has two important properties. (a) In each clustering pass through the resemblance matrix all local minima (or maxima) are located. After this the mutually closest objects or clusters are united to form new clusters. This general strategy is called L-algorithm in this paper. (b) In cases when there is no unique smallest (or highest) value in any column of the matrix, the algorithm may group together more than just pairs of objects ('multiple fusion'). It is worthwhile to consider the following reasoning. Let we have three objects denoted by z_1, z_2 and z_3. According to Sibson (1971) let the similarity matrix be specified by

	z_1	z_2	z_3
z_1	1	λ	$\lambda-\varepsilon$
z_2		1	λ
z_3			1

We can see at once that the first level at which any group can be formed is λ, and, if the conventional algorithm is used, we ought to take an arbitrary choice between pairs z_1z_2 and z_2z_3. But, instead of this, we fuse all objects at hierarchic level determined by the nature of the strategy used. For example, in the case of group average sorting, the hierarchic level will be $\lambda-\varepsilon/2$. However, two disadvantages arise, the chain effect may slightly increase and reversals may occur in the dendrogram obtained not only with the centroid methods

but also with the complete linkage method.

When the classification is based on binary data the arbitrary choices may frequently make the results inconsistent. Let us consider two classifications (Figs 2.13.2-3.) of the same set of phytosociological quadrats from rocky grassland communities (Podani 1978b). Both dendrograms were obtained with the same algorithm namely the edge-density optimization strategy (see below). In two cases which were similar to the hypothetical situation discussed above, random choices were made. In the first run the algorithm decided on the following pairs: 2-5 and 28-32. On the contrary, in the second run the algorithm joined pairs 4-5 and 32-50. The considerable differences between the two resultant dendrograms are due only to these arbitrary choices. It is obvious that none of the two dendrograms can be regarded as a result of an 'objective classification'. Therefore it was preferable to use the well-defined algorithm described above which performed two multiple fusions and produced fairly unambiguous result (Fig. 2.13.5.).

(2) The original data may be read from cards, magnetic tape or disk. Parameter IFILE is to be set equal to 0 for input from cards. Otherwise IFILE is equal to the data set identifier. It must be borne in mind that the use of the following identification numbers is prohibited for technical reasons.

in program NCLAS	9, 11
in program HMCL	9
in program INFCL	9.

The number of objects (M) to be classified must not exceed a certain limit in each program. The number of attributes in programs INFCL is also limited, but in the other two programs it is limited only by the magnetic storage capacity. These limits can be readily changed for other computers than CDC 3300, if desired, by redimensioning most of the vectors and arrays in the programs and the subroutines. The required dimensions are given in Tables 3.2.2-3. and also in the listings.

The input format is provided by the user on a separate card with the following restrictions. In programs NCLAS and HMCL F specification can be used only even if the data are represented in integer format. On the contrary, program INFCL can read integers only using I specification. The input data are written on disk file therefore all of the three programs may carry out several analyses using the same data set.

(3) The output is controlled by parameter IMODE. When IMODE is set equal to 0, the programs print title, run parameters and the input format; list the serial number of the clustering cycles and the fusions, the identification number of groups amalgamated, the number of objects previously united to form each of these groups, the value of the resemblance coefficient or the change in heterogeneity and the value of heterogeneity for the newly formed cluster. It is well-illustrated by the self-explanatory printout examples in the appendix. It is to be noted that new groups are identified by the smallest identification number of the group (i.e., if objects 5 and 10 are joined, then the new group will be identified by number 5).

All of the agglomerative programs utilize subroutine DEND to convert the results to a dendrogram which is drawn on the line printer. The dendrogram may take up several pages depending on the number of classified entities. In order to show slight changes in the hierarchy, the scale against which the dendrogram is plotted is obtained according to the lowest and the highest criterion values.

If IMODE is set equal to 2, the programs, besides printout, punch a sequence list of the classified entities in FORMAT (20I4).

Special characteristics and underlying mathematical principles of each agglomerative programs are given below.

Program NCLAS performs route-optimizing (r-hierarchical) clustering based on the combinatorial equation of Lance and Williams (1966). The user has a choice between 23 measures of

resemblance and eight types of algorithm using options MCOE and MSO, respectively. The input data may be transformed in three ways, standardization by standard deviation, standardization by range and logarithmic transformation. The type of standardization is controlled by parameter ISN. The standardization does not destroy the original data stored on disk file.

It is possible to analyze the same resemblance matrix with several algorithms without repeating computations. In the first run parameter MSTD is to be set equal to 2, and then, in the subsequent runs, MCOE is to be set equal to 0. There is, however, an exception. This opportunity can not be used if MCOE= =8 (see the compatibility matrix for four run parameters in Table 3.3.1.).

The resemblance function used here are well-known ones with two exceptions. The author proposes the application of the Euclidean distance to mixed data on the analogy of Gower's similarity coefficient, since the latter metric can not be applied to the sum of squares agglomeration. The Euclidean distance for mixed data is defined by

$$t_{jk} = \sqrt{\sum_i w_{ijk} \left(\frac{x_{ij} - x_{ik}}{q_{ijk}} \right)^2}$$

where x_{ij} and x_{ik} are the scores of objects z_j and z_k, resp., for attribute a_i. The scores q_{ijk} are assigned as follows

a. For binary attributes $q_{ijk} = 1$ for all j and k.

b. For disordered multistate attributes

$$q_{ijk} \begin{cases} x_{ij} - x_{ik}, & \text{if } x_{ij} \neq x_{ik} \\ 1, & \text{if } x_{ij} = x_{ik} \end{cases}$$

c. For ordered multistate and quantitative attributes

$$q_{ijk} = \max_j \{x_{ij}\} - \min_j \{x_{ij}\}, \text{ for all } j \text{ and } k$$

i.e. q_{ijk} is the range of attribute i.

The weight w_{ijk} is set equal to 1 if the comparison of z_j and z_k is valid for attribute a_i. When this comparison is not possible because of unknown data, w_{ijk} is set to be equal to 0. However, in the latter case, the axiom of the triangle inaequality is not satisfied, and this function will not be metric.

The weighted dissimilarity index (2.8.18.) was proposed by the author (Podani 1978a) for phytosociological data. When this function is used, the attributes are weighted according to the probability of their presence in the entire set analyzed.

If Gower's coefficient of similarity or the Euclidean distance for mixed data is used (i.e.,if MCOE.GE.41) , we must pay attention to the following restrictions

a. The number of attributes must not be greater than 999.

b. Unknown data is to be coded by negative values in the input matrix. Consequently, for known data only positive or zero values can be used.

c. The attributes must be arranged in a strictly defined sequence in the data matrix as

binary attributes
non-ordered multistate attributes
ordered multistate and quantitative attributes

The absence of any type of these attributes is of course allowed.

d. The data should be represented in constant format using F specification without taking the type of data into consideration.

HMCL is a homogeneity-optimizing (h-hierarchical) classification program. The basic combinatorial equation (2.16.1.) was developed by the author (1979b). The user has a choice between three clustering strategies: optimization of dispersion, average dispersion or the edge-density of the new clusters. The first two methods are closely related to the sum of squares agglomeration but those pairs of groups are joined at which the dispersion (or the average dispersion) of the new group is minimum. Euclidean and chord distance may be applied to measure the heterogeneity of groups.

The edge-density optimization technique (Podani 1979b) can be used in the analysis of binary data. The homogeneity of the groups may be defined in two alternative ways. If negative matches are considered valid, the edge density is the generalization of the simple matching coefficient (2.8.11.) for groups consisting of more than two elements. If negative matches are excluded, we have the generalization of the Russell - Rao coefficient (2.8.13.).

The compatibility of parameters determining the type of sorting algorithm (MSO), the resemblance function used (MCOE) and the way of standardization (ISN) is summed up in Table 3.4.1.

Program INFCL classifies the entities on the basis of information theory. All of the four techniques programmed are applicable to binary data only. The program accepts integers of any kind and transforms them to binary form according to

$$x'_{ij} \begin{cases} 1, & \text{if } x_{ij} \geq 1 \\ 0, & \text{if } x_{ij} < 1 \end{cases}$$

The classification may be based on two criteria, the mutual information between attributes and the preferential information heterogeneity (PIH) of groups. The latter measure is known

under the commonly used and rather ambiguous term of 'information content'. It can be easily shown that, in fact, the 'true' information content of a population is a quite different quantity being a diversity measure (see below). The term of preferential information heterogeneity was firstly used by the author (Podani 1978b) to clear up this question. The PIH of a group Z_h can be calculated using the formula given by

$$m\hat{I}_{Z_h} = n\, m \log m - \sum_i (k_i \log k_i + (m-k_i) \log (m-k_i))$$

where m is the number of elements in Z_h, k_i is the frequency of the presence of attribute a_i in Z_h, n is the number of attributes describing Z_h. The quantity $m\hat{I}_{Z_h}$ is the overall measure of the preference of attributes for certain objects against the rest (cf. Juhász Nagy 1972). It is easy to see, that the preference of attribute a_i is greatest if its entropy

$$m\hat{H}_i = m \log m - k_i \log k_i - (m-k_i) \log (m-k_i)$$

is maximum. On the other hand , if the attribute i does not prefer any object (i.e., a_i is present in or absent from all objects in Z_h), its entropy is zero.

The PIH of a set is usually greater than the 'true' information content (TIC) because of the mutual information contained in the set of attributes. In the case of synbiological entities, e.g. zooplankton samples or phytosociological quadrats, the 'true' information content is called floral or faunal diversity (see Juhász Nagy 1972, Juhász Nagy et al. 1973). In general, TIC is the measure of the diversity of any set described by binary attributes. This quantity is maximum, if there are no two entities in the set which agree in all attributes (i.e., any two entities differ from each other in one attribute at least). The TIC can be calculated according to

$$m\hat{I}D_{Z_h} = m \log m - \sum_{g=1}^{\omega} k_g \log k_g$$

where ω is the number of the possible attribute combinations including the full zero combination as

$$\omega = 2^n$$

and k_g is the frequency of the gth attribute combination.

As mentioned above, the mutual information contained in the attributes can be simply derived by extracting TIC from PIH.

The two heterogeneity measures can be applied to either the r-hierarchical or the h-hierarchical algorithm. The heterogeneity measure and the sorting method may be chosen using parameter MSO.

ASSINF, the fourth program of the SYN-TAX package has to be dealt with separately since it has very different characteristics from the agglomerative programs. This is a classification program using a divisive, monothetic method. The set or subset of entities is subdivided in each step of the analysis according to that attribute which has the maximum sum of mutual information calculated with all the others (Podani 1979a). This method is a slight modification of the association analysis worked out by Williams, Lambert and Lance (see Lance - Williams 1968).

This program, in contrast with the agglomerative programs, can carry out only one analysis in a single run. The data are read in integer form and transformed to binary data as done by program INFCL. Since a disk file used by the program is identified by number 10, the parameter IFILE must not be set equal to 10. The array and vector dimension statements may be redefined according to either Table 3.6.1. or the documentation at the beginning of the list of the program and associated subroutines.

Labels for attributes (usually plant species) are read separately prior to data input. We have eight characters to code or abbreviate the name of each attribute. Then one card contains labels for ten attributes (see the input example of this program in the appendix). We must be careful that the sequence of attributes should be identical in the label cards and in the data matrix as well.

Automatic stopping rule is not adapted to the algorithm. The program subdivides only those groups which consist of more than K elements. The appropriate value of K is determined by the user, but K must not be less than 2. When more than one divisive attribute is obtained, their distributions are tested. The subdivision is ignored in ambiguous cases.

The printout list of this program is self-explanatory (see the appendix).

(Received December 1979, Revised March 1980)

Note for IBM users. More recently the SYN-TAX programs are adapted to an IBM 3031, System 370. All programs are re-dimensioned such that programs NCLAS, HMCL and ASSINF may be used up to 300 objects, programs INFCL is re-dimensioned for 150 objects. Slight modifications and implementations were made throughout. Listings are available upon request from the following address

Dr. J. Podani
Research Institute for Botany
Hungarian Academy of Sciences
VÁCRÁTÓT
2163 Hungary