COENOSES 3 (2): 111-119, 1988



# SYN-TAX III. A PACKAGE OF PROGRAMS FOR DATA ANALYSIS IN COMMUNITY ECOLOGY AND SYSTEMATICS

J. Podani
Research Institute of Ecology and Botany, Hungarian Academy of Sciences, Vácrátót, H-2163 and Department of Plant Taxonomy and Ecology, L. Eötvös University, Budapest, Kun B. tér 2, H-1083

**Keywords:** Clustering, Comparison of results, Computer programs, Ordination, Pattern analysis, Ranking, Simulated sampling

**Abstract:** This new release of the package contains twenty-two FORTRAN programs written for both mainframe computers and IBM/PC AT and XT compatible machines. Widely used standard multivariate methods and specific data analytical techniques, some of them suggested by the author, are represented in the package. The procedures programmed include hierarchical and nonhierarchical clustering, ordination, character ranking, comparison of classifications and ordinations, generation of consenus partitions and ordinations, Monte Carlo simulation of the distribution of coefficients of partition agreement, simulated sampling based on digitized point patterns, and information theory functions for the analysis of species assemblages. This paper provides general information on the programs, detailed documentation is given in the user's manual.

## Introduction

The first version of the SYN-TAX package (Podani 1980) contained four programs for hierarchical clustering, three included agglomerative methods, whereas the fourth was designed to perform an information theory variant of association analysis. The second release (Podani 1984c) represented a significant extension of the first; in addition to the classification methods, many other facilities of data analysis, such as ordination, comparison of dendrograms, simulated sampling, and character ranking were included. The programs have been used in a wide range of biological applications providing continuous encouragement to improve the existing programs and to extend the scope of the package. The increased availability of personal computers was another stimulation to develop SYN-TAX III.

As the previous versions, SYN-TAX III includes some methods also available in other well-known packages (e.g., ORDIFLEX, Gauch 1977; CLUSTAN, Wishart 1975). For example, NCLAS2 offers options for several widely used hierarchical agglomerative techniques; likewise, ordination methods have also been programmed by many authors. However, even these programs have some unique features (e.g., option for single link and suboptimal fusions to resolve ties in agglomerative clustering, lineprinter graphics, etc.); and the procedures never programmed elsewhere (e.g., in programs MINGFC, PARREL, HMCL2, DENDAT, SAMPROC, and INPRO3) make SYN-TAX III quite distinct from all other software currently available for taxonomists and ecologists.

Each program has two variants, the first prepared for mainframe computers in IBM's standard FORTRAN IV and the second written in Micrsoft FORTRAN 77, version 4.0, for IBM XT, AT and compatibles. These alternatives produce practically identical results, but there are some minor differences in printouts, file handling, etc. Upon request, the package will be modified to any other installation with at least 254 K RAM.

The programs may be logically arranged into five groups: 1) cluster analysis; 2) ordination; 3) evaluation and comparison of ordinations and classifications; 4) simulated sampling and analysis of synphenetic pattern; and 5) character ranking. Of course, the distinction between these categories is not sharp. For example, program MINGFC has been designed to construct a hierarchical set of consensus partitions, but the clustering algorithm may also be used to analyze distance matrices calculated by other programs. Further relationships are established by jointly used subroutines and chaining (the output saved by one program is potential input for others in many cases).

First, a general description of the SYN-TAX III programs is presented. Then, basic information on each program is provided following the five-group classification of the programs. Detailed input-output specifications and examples are not discussed here, these are included in the user's manual (Podani 1988c).

## Features of SYN-TAX III

The programs have many standardized features for the convenience of the user. These are collected here so that the discussion of programs may be confined to their special characteristics.

### *Modes of operation*

The programs may be run in two different modes. In

interactive mode the parameters and options must be specified during program execution. A menu is displayed on the screen whenever the user is prompted for reply. In this mode only the input data, input format, and - in some programs - labels for variables are read from files. The interactive mode is recommended for beginners and is in fact the most convenient way of specifying input/output options if a single analysis is done only. However, if one wishes to use several combinations of parameters within the same run of the program (this is allowed in several programs), the batch or noninteractive mode becomes more straighforward. Also, batch execution is a convenient way to run jobs while away from the computer. In this case almost all parameters and options must be collected in a separate file and the user is asked only a few questions regarding mode and files at the start of the analysis.

As far as basic computations are concerned, there are no differences between the two modes. Some programs, however, offer more facilities in the interactive mode. For instance, in program PRANA the scattergram of transformed ordinations may be printed only in the interactive mode because the option whether or not to print the scattergram would be complicated to specify in advance.

*Input data and format*

All but one SYN-TAX III programs, the exception being PARTEST, require input data in form of either a raw data matrix, distance matrix or some special arrays (containing descriptors of classifications and ordinations, coordinates, etc.). The raw data matrices must be full, with variables in rows (except in program PRINCOMP). Condensed forms are not supported. Missing entries in the data matrices are usually not allowed because they are treated as zeros. If special resemblance functions are used which are applicable to missing values, each missing entry is to be indicated by a dummy value preceded by a negative sign. If a distance matrix is input, only the upper semimatrix and the diagonal are required and are read by columns into a one dimensional array. The SYN-TAX III programs cannot analyze asymmetric matrices.

The free input format for data has to be provided by the user on the first record of the data file in most programs. Usually real field specification must be used even if integers are read. Some programs require that data be prepared in a fixed input format.

*Files*

The mainframe (MF) versions use device 1 for input data and device 2 for parameters. Files 3, 4 and 7 are reserved for temporary data storage and output arrays which can be analyzed by other programs (see next section). There is an option to save printed output on file 8 instead of sending it to the display. These input/output devices must be specified prior to execution according to the rules of the actual operation system.

For personal computer (PC) versions, the filenames and extensions must be specified runtime in both interactive and batch modes. Entering filenames may be avoided using the default names listed in the menu. Save files with default names must be renamed after program execution to prevent overwriting them in the next run of any SYN-TAX program. Thus, saved results may be retrieved when required (see below).

*Chaining*

The SYN-TAX save files contain intermediate or final results that can be used as input for other programs. The agglomerative clustering programs save dendrogram merge matrices for subsequent dendrogram comparisons. All hierarchical classificatory programs may be instructed to output the sequence of objects in the dendrograms for use in analysis of concentration (BLOCK) and comparison of partitions (PARCOM2). The agglomerative programs prepare and save merge matrices for dendrogram comparisons by programs DENCOM and DENDAT. The distance matrices calculated and saved by a SYN-TAX III program may be further analyzed by other programs. Ordination scores are written onto files for Procrustes analysis to compare ordinations or to construct a consensus ordination (PRANA). Sampling simulator programs (ELSAM2, SAMPROC) output raw data matrices for clustering, ordination and pattern analysis. Finally, the ranking procedures save the rearranged data matrix so that multivariate analyses based only on highly ranked variables become possible.

*Lineprinter graphics*

Many different graphics libraries exist, but they may not always be available. Therefore, to ensure universal applicability of the package, plotting procedures are not included. Instead, dendrograms and scattergrams are produced by the lineprinter. They appear in the same form independently of the type of printers used. The dendrograms, as output by five programs, may take up several pages if many objects are classified. In order to show fine details of the hierarchy, the full dendrogram is not printed. All fusion levels are shown but edges below the lowest actual level are cut. The number of characters determining height of the dendrograms is specified by the user (72 is recommended for output on the screen, 120 for printers). Ordination scattergrams are produced by four programs. The configuration of points is shown in two dimensions selected by the user. Objects are identified by serial numbers on the ordination plane, the maximum number of objects is 999. To keep the number of overlapping, and therefore omitted, serial numbers to the minimum, only a reduced or-

dination space is portrayed which is determined on the basis of the actual extreme values on the axes. The height of the ordination plane is 60 characters, the width is specified by the user (recommended values are 76 for display and 120 for printer) except in program BLOCK where the fixed value of 76 is large enough to show the maximum of 20 points (releve and species groups). Program ELSAM2 prints a map showing the location of plot centres in simulated sampling.

*System requirements, dimensions, problem sizes*

The MF versions are dimensioned so that they can be run on an IBM 370 under the CMS operation system with 512 K virtual memory. For many of the PC programs two executable codes have been made, one for IBM XT compatibles with 360 K RAM and the other for AT machines with 520-640 K. For example, the "smaller" version of program NCLAS2 classifies up to 200 objects, whereas the "large" version can handle as many as 420 objects. In many programs 360 K is large enough for any practical poblem size thus allowing all machines to use the same executable code. It is assumed that the operation system in MS-DOS 2.1 or higher.

Maximum problem size does not always depend on limitations of memory. For example, the number of variables (i.e., rows of the data matrix) is unrestricted in programs NCLAS2, HMCL2 and PRINCOOR; the upper limit is determined by the maximum file size allowed by the operation system. In case of mixed data, however, some limitations hold (namely the maximum is 999). In program PRINCOMP the number of observations may not exceed 999 because this is the maximum number of points that can be displayed in scattergrams.

## Cluster analysis

In the SYN-TAX III package major emphasis is placed on classification; seven programs are devoted to cluster analysis. Four of them (NCLAS2, HMCL2, INFCL2, and MINGFC) include agglomerative algorithms, whereas program ASSIN2 is divisive. Program PARREL has been designed to improve output or random partitions iteratively according to a global optimality criterion. Program MINSPAN, written for constructing minimum spanning trees, is also included in this category because of its obvious relationship to the clustering methods.

The agglomerative programs share several features. Specific routines are included to solve fusion ambiguities often encountered during the clustering process especially if binary data are analyzed. Packages currently available resolve ties by arbitrary choices and the user is not even warned that the fusion sequence is not unique. Spurious details that are potentially introduced in this way into the analysis may have serious impact on the results as demonstrated by Podani (1980). The agglomerative programs have two options for tie-breaking: 1) tied pairs are fused according to a single linkage criterion, and 2) the fusion of tied objects is omitted and a suboptimal fusion with which no ties are associated is done instead (Podani 1988 describes the details). In addition, programs NCLAS2, HMCL2 and INFCL2 may be run in the "traditional" manner, i.e., disregarding ties by arbitrary fusions.

I. Program NCLAS2

Distance-optimizing hierarchical classification is performed. Eight methods compatible with the Lance-Williams (1966) recurrence equation may be selected: 1) single linkage, 2) complete linkage, 3) average linkage, 4) centroid, 5) beta-flexible (Lance-Williams 1967), 6) median, 7) simple average, and 8) beta-gamma flexible (DuBien-Warde 1979). The program provides 29 distance coefficients, some of them being dissimilarity or similarity measures expressed in distance form. In interactive mode squared distances or dissimilarities may also be used. Two functions can be applied to data sets with mixed variable types and missing values. Prior to calculating the distances, five data standardization and transformation procedures may be chosen.

The algorithm fuses reciprocal nearest neighbors (RNN) in each cycle of the clustering process for methods 1, 2, 3, 5, and 7. These strategies satisfy the reducibility condition (Bruynooghe 1978, Gordon 1987) allowing such an acceleration of the analysis. However, only a single fusion (of the closest pairs) is done in methods 4, 6 and 8 because the above condition does not hold for these strategies.

II. Program HMCL2

It performs homogeneity-optimizing hierarchical classification. The basic combinatorial equation (Podani 1979, 1988b) is compatible with the following eight strategies:

1) Minimization of average distance within new clusters (Anderberg 1973, Podani 1979);
2) Minimization of sum of squares within new clusters (Anderberg 1973);
3) Minimization of variance in new clusters (Anderberg 1973);
4) Minimization of the increase of weighted average distance (Podani 1988b);
5) Minimization of the increase of unweighted average distance (Podani 1988b);
6) Minimization of the increase of sum of squares (Ward 1963, Wishart 1969, Orloci 1978);
7) Minimization of the increase of variance (Diday et al. 1982); and
8) Lambda flexible (Podani 1988b).

The distance coefficients and standardization procedures agree with those offered by program NCLAS2. The user has a choice whether reciprocal nearest neigh-

bors or only the closest pairs are fused in each cycle of the analysis; for methods 2, 3 and 6 the RNN algorithm is straightforward. For the other methods the reducibility condition does not hold and the alternative algorithms may produce quite different results.

III. Program INFCL2

Information statistics serve as descriptors of cluster homogeneity in agglomerative clustering. Unlike in the preceding two programs, the solution of information theory clustering is not combinatorial, so the whole data set must be retained in computer core throughout the calculations. Only binary data are analyzed, other scores are automatically transformed to binary form. The methods are as follows:
1) Minimization of the pooled entropy of variables (preferential information heterogeneity or local distinctiveness, cf. Juhász-Nagy - Podani 1983; often termed as 'information content') in the new clusters;
2) Minimization of the increase of pooled entropy ('information analysis', Williams et al., 1966, Orloci 1969);
3) Minimization of the mutual information of variables in the new clusters (Podani 1980); and
4) Minimization of the increase of mutual information among variables (Orloci 1969).

The fusion algorithm uses either reciprocal nearest neighbors or closest pairs. It seems, however, that all the four methods satisfy the reducibility condition (although a proof is needed) so that the first algorithm is preferable.

IV. Program MINGFC

Its underlying principles radically differ from those of the other three agglomerative methods. The fusion of two objects is conditioned upon the 'goodness' of the whole classification rather than upon pairwise measures (Podani 1988a). The optimality criterion is relatively simple: the average of within-cluster distances divided by the average of between-cluster distances. Two objects are amalgamated if the increase of this ratio upon their fusion is the minimum. The final result is an incomplete 'dendrogram' with the last fusion level missing (for this the ratio is undefined). The program contains no routines for calculating distance matrices; they can be prepared in advance by program NCLAS2. In addition to general purpose clustering, the method can be used to generate a series of nested consensus partitions for a set of k different partitions of the same objects (Podani 1988a).

V. Program ASSIN2

Monothetic divisive clustering ('association analysis') is performed using one of two information theory division criteria:
1) Sum of mutual information between species pairs (Podani 1979a); and
2) Information fall (i.e., decrease of pooled entropy, Lance-Williams 1968).

The analyses are based on binary data even if quantitative scores are entered. During calculations the whole data set is stored in memory. There is no stopping rule applied; the subdivision of a group stops if its size is not greater than a specified threshold or if two or more variables are found and their distributions conflict. Since the small groups are not split into individual objects, no merge matrices are saved and the printed dendrogram differs from those produced by programs I-IV.

VI. Program PARREL

Objects in a starting partition are relocated iteratively to minimize the ratio of average within-cluster and average of between-cluster distances. The starting partition is either randomly generated or derived from hierarchical classifications. In the general situation, program PARREL improves a partition of objects based on any distance matrices calculated by other SYN-TAX III programs. A special application is the improvement of suboptimal consensus partitions generated by program MINGFC.

VII. Program MINSPAN.

This program constructs a minimum spanning tree from distance matrices. The options for distance function and data transformation are those offered by programs NCLAS2 and HMCL2. In each step of the computations two points are connected if they belong to separate subgraphs and their distance is the minimum (cf. Sneath-Sokal 1973). The output contains a list of these steps, but no graphical result is printed. The user is advised to run program PRINCOOR for principal coordinates analysis with the same options and then the tree may be easily drawn on the resulting scattergram of the first two dimensions. This program is included among the classificatory methods since the successive removal of the longest edges from the graph implies a divisive counterpart of nearest neighbor clustering (cf. Pielou 1984).

**Ordination**

The purpose of ordination is to evaluate continuous variation in the data by reducing dimensionality (see e.g., Orloci 1978, Pielou 1984, Digby-Kempton 1987, for an exhaustive treatment of the topic). Although only two ordination programs have been included, they offer a wide variety of metric scaling procedures. All analyses are based on the eigenanalysis of symmetric real matrices. The resulting scattergrams are either displayed on the screen or sent to the printer. The output of starting matrices, eigenvectors and percentage contributions is optional. Ordination scores are saved

for subsequent evaluation by clustering or Procrustes analysis.

VIII. Program PRINCOMP

Its basic strategy is principal components analysis from matrices of cross products, covariances or correlations (noncentered PCA, centered PCA, and standardized PCA, respectively). Furthermore, three types of correspondence analysis (CA), particularly useful to analyze categorical data, may be performed: 1) raw coordinates are weighted averages of column coordinates, 2) symmetric weighting, and 3) column coordinates are weighted averages of row coordinates. The lower the eigenvalues the greater the differences among the results of alternative analyses. For details of the underlying principles of PCA and CA see Legendre-Legendre (1983), Pielou (1984) and Greenacre (1984).

IX. Program PRINCOOR

The classical solution to the metric multidimensional scaling problem (cf. Mardia et al., 1979) is provided, the method is also known as principal coordinates analysis (Gower 1966). Options for distances and standardizations are the same as in programs NCLAS2, HMCL2, and MINSPAN. The analysis may start from raw data or directly from distance matrices. If the input distances do not satisfy the metric axioms, the results may also be interpreted with certain limitations (see e.g., Digby-Kempton 1987, for the treatment of negative eigenvalues).

**Evalutation of classifications and ordinations**

The numerous clustering procedures usually provide different solutions for the same set of objects, unless the group structure is obvious. Similarly, scaling procedures may also yield diverging results. In addition to choices regarding a particular method, the selection of sampling design, variables and data types used, etc., is also important in influencing classifications and ordinations. This problem calls for comparisons to reveal the relative impact of our decisions upon the results. Comparisons may also be necessary when temporal effects (e.g., in succession studies) are to be indicated.

Besides comparisons, the alternative results may be synthesized by consensus methods so that the consensus classification or ordination reflects agreements and disagreements among the competing results. When variables and objects are simultaneously classified, analysis of concentration is recommended to assess block structure in the rearranged data matrix.

The SYN-TAX III package is intended to support such purposes. In addition to program MINGFC, already discussed among the clustering programs, six other programs contain procedures for the evaluation of results of multivariate analyses.

X. Program BLOCK

This program prints rearranged binary data matrices (e.g., phytosociological tables) and performs analysis of concentration (Feoli-Orloci 1979). The classification of objects and variables must be done in advance using any method of data analysis. For example, species and relèves may be classified in two runs of program NCLAS2 which provides a revised sequence of the clustered items. The sequence vectors and the original, non-ordered data matrix represent input for program BLOCK. The printed output includes chi $^2$ statistics, canonical correlations and results of a symmetric CA of row and column groups of the rearranged data table. Ordination scattergrams show the configuration of points in the first two dimensions.

XI. Program DENCOM

Many coefficients of dendrogram dissimilarity are computed: Euclidean distance and absolute differences using 1) topological differences, 2) cophenetic differences, and 3) cluster membership divergences (for details, see Podani-Dickinson 1984). Further measures are: 4) number of object triplets for which the ultrametric relationships conflict in the two dendrograms (Dobson 1975) and its normalized form, 5) mismatched edge difference (Podani 1982), 6) absolute edge difference (Robinson-Foulds 1979), 7) edge matching coefficient (suggested by Podani 1982 after Robinson-Foulds 1979). Since the calculation of function (4) is tedious for large dendrograms, its computation may be rejected.

Two dendrograms are compared in a single run. The dendrograms are described by two merge matrices. Each matrix specifies the fusion sequence of objects and clusters in terms of m-1 rows, where m is the number of objects classified. There are five columns in the matrix, the first two contain cluster identifiers, the second two indicate cluster sizes, and the fifth contains the hierarchical levels. The identifier of a cluster is always the smallest serial number of its component objects. SYN-TAX III programs I-IV prepare the merge matrices of dendrograms in this way.

XII. Program DENDAT

This program has been written in order to prepare data for the multivariate analysis of several dendrograms that represent hierarchical classification of the same objects. Any combination of 5 characteristics of dendrogram structure may be incorporated in the analysis. These are topological difference, cophenetic difference, cluster membership divergence, subtree membership divergence, and partition membership divergence (Podani-Dickinson 1984). The descriptors are standardized so that they equally contribute to the distances between dendrograms. Normally a Euclidean distance matrix is output, but if other function of den-

drogram dissimilarity is sought, a data matrix of dendrogram descriptors may result. The input is a sequence of merge matrices prepared as described above.

XIII. Program PARCOM2

This program compares two partitions using twelwe distance and dissimilarity coefficients: 1) chi $^2$, 2) 1-Cramer (Anderberg 1973), 3) mutual predictive power, 4) predictive power of either partition to the other (both by Goodman-Kruskal 1954), 5) 1-simple match (Rand 1971), 6) Euclidean distance, 7) 1-Ochiai (Fowlkes-Mallows 1980), 8) 1-Jaccard, (Downton-Brennan 1980), 9) 1-Sorensen ('percentage difference', Gauch-Whittaker 1981), 10) minimum number of transfers, mergences and divisions needed to transform one partition to the other (Day 1981, Gordon 1980, 1981), 11) normalized form of 10 ('misclassification' index, Podani 1986), and 12) Day's sigma (Day 1981). For measures 5-9 and 12, possible maxima and normalizations are also calculated (see Podani 1986 for details of normalization). Coefficients 10-11 are not computed if the number of clusters exceeds six.

XIV. Program PARTEST

This program performs Monte Carlo simulation of distributions of distances between partitions of the same set of objects. The number of clusters is assumed to be the same in both classifications; the maximum is 6. The distributions are either simply restricted (only the number of groups has to be specified) or doubly resctricted (number of groups plus group sizes are to be specified). Sampling distributions of five coefficients are generated in a single run. These are: 1) one-complement of mutual predictive power, 2) minimum number of transfers, mergences and divisions needed to transform one partition to the other, 3) 1-Ochiai, 4) 1-Jaccard, 5) 1-Rand. There is an option for the normalization of the latter three measures, but in that case another run of the program is necessary. The output includes frequency distributions, empirical probability distributions, cumulative probability distributions and sample statistics (mean, variance, skewness, and kurtosis). In addition, the program may be instructed to simulate the distribution of minimum distances among k randomly selected partitions to perform multiple comparisons.

The simulated distributions may be used to test the significance of distance measures. Possibilities are reviewed and examples from vegetation science are provided by Podani (1986).

XV. Program PRANA

Different ordinations (configurations) of the same points are evaluated. Two main types of assessment must be distinguished: 1) comparison of ordinations by Procrustes analysis (Schoenemann-Carroll 1970, Gower 1971, Sibson 1978), and 2) construction of a consensus ordination for $k$ ordinations via generalized Procrustes analysis (Gower 1975, see also Digby-Kempton 1987). In the first case pairwise comparisons of several ordinations may be achieved in a single run of the program and the resulting distance matrix is saved. To obtain a symmetric distance measure between ordinations, the user has an option to rescale every ordination to unit sum of squares. The generalized Procrustes analysis is an iterative procedure during which further rescaling can be made (it is recommended). The number of points must exceed the number of dimensions (this condition usually satisfies). The output includes coordinates and scattergrams for rotated and/or consensus configurations.

## Simulated sampling and analysis of synphenetic pattern

Computer simulated sampling provides an efficient tool to evaluate the appropriateness of particular sampling designs for estimation or typification purposes and to analyze spatial point patterns (Podani 1987 reviews some possibilities with examples). The SYN-TAX III package contains two sampling simulator programs (ELSAM2, SAMPROC), another program for both sampling and subsequent analysis of inter-plot resemblance (EXPRES), and a fourth one (INPRO3) designed for the computation of information theory characteristic functions for pattern analysis.

The sampling simulators are applicable to digitized point patterns, i.e., the actual extension of individuals is assumed to be point-like. The input for these programs is a data matrix describing the position of plant individuals in terms of Cartesian coordinates on the plane of a rectangular study area. The plots are placed by the program within this region because they cannot overlap the boundary line.

XVI. Program ELSAM2

This program simulates random plots of circular, rectangular or elliptical shape with random or uniform orientation. Being independently located, the plots may overlap with one another. The total overlapping area increases as sample and/or plot size increases, and this fact mut be considered when evaluating the results. For estimation purposes the overlap of plots causes no problems, but signifance tests must be done with caution based on data produced by this program. The output matrix is a species by plots table written onto disk file. The location of plot centres within the sampled area is shown by a scattergram, facilitating interpretation of results that are based on samples taken by program ELSAM2.

XVII. Program SAMPROC

A flexible sampling design, which includes systematic, restricted random and fully random sampling (Podani 1984a) is used to obtain samples with different arrangements of plots. Only circular or rectangular plots can be used; the blocks (compartments) within which the plots are placed are always square. The size of blocks is specified by the user; sample size is automatically determined. If the size of contiguous blocks is defined so that the grid does not cover completely the study region, a narrow strip on the upper and/or right side of the area will be excluded from sampling.

In the simplest case, the program performs systematic sampling. The study area is subdivided into blocks with size specified in advance, and a single plot is located at the centre of each block. If a sampling process is carried out, the sequence of increasing block sizes must contain the size of contiguous blocks. Full randomization is achieved if the side length of largest blocks is at least twice as long as the longer side of the study region. In the sampling process the plots are always located randomly within the block (except in systematic sampling as mentioned above). The resulting series of data matrices is saved on file.

XVIII. Program EXPRES

This program simulates random and nonoverlapping pairs of plots and calculates the expectation of six resemblance coefficients in turn. Plot shape is either circular or rectangular. The expectation is estimated as the average resemblance between random pairs of plots (Podani 1984b) which is indicative of synphenetic pattern and serves as an alternative to information theory characteristic functions (see next section). The program computes expected values and standard deviations for the Sorensen index, Russell-Rao index, simple matching coefficient, binary distance, Euclidean distance per unit area, and chord distance. No data are output.

XIX. Program INPRO3

This is an expanded version of program INPRO described in Juhász-Nagy - Podani (1983). Information theory functions to evaluate autophenetic and synphenetic pattern are calculated from a species by plots data matrix. These functions are based on Shannon's entropy function and may be used to the evaluation of the consequence of sample plot size and for the study of succession. The program accepts any kind of data, the scores are automatically converted to binary form.

Functions informative on the synphenetic pattern are: florula diversity, florula evenness, local distinctiveness, associatum, number of realized species combinations (Juhász-Nagy - Podani 1983), and total dissociatum and entropy functions for local valences, floral valences, local invalences and floral invalences (Juhász-Nagy 1984, 1985). In addition, the entropy, associativity, and dissociativity for each species are printed together with the corresponding values of subdiversity and subassociatum (see Juhász-Nagy 1984, 1985). The printing of species combinations is optional. Several data sets may be analyzed in the same run of the program, allowing rapid evaluation of plot size effects.

**Character ranking**

Some character ranking procedures are alternatives to ordination methods when reduction of dimensionality in the data is the objective. Whereas ordinations introduce artificial dimensions that are not always interpretable, character ranking keeps the original variables by creating their order of importance. Of course, the highly ranked variables are usually less efficient than the ordination axes. Other ranking procedures do not remove the effect of already ranked characters, these procedures are recommended to select unimportant variables prior to clustering.

The ranking programs output a data matrix in which the variables are arranged according to their rank order. Variables not ranked are collected at the end of the new data file. The rearranged table allows the user to apply multivariate techniques to reduced data sets and then to assess how the results are influenced when less important variables are successively omitted (see Podani 1984a).

Three ranking programs are included in the package. All of them accept labels to designate variables.

XX. Program RANSQ

The variables are ranked according to one of the following dispersion criteria: 1) cross products, 2) covariance, and 3) correlation. From the dispersion matrix the rank order is obtained either by eliminating the residuals (Orloci 1973) or simply by calculating the contribution of each variable to the total variation (i.e., residuals are not eliminated).

XXI. Program RANKINF

This program is applicable to ranking binary characters based on Orloci's (1976) information criterion. The computations are slow for very large data sets, therefore the user may specify a percentage threshold beyond which the analysis stops.

XXII. Program SUMR

This is another program for ranking binary characters. In this case the criterion is the same as the one used in association analysis (Podani 1979a), namely, the sum of mutual information values. Two strategies may

be chosen depending on whether or not the characters selected in the previous cycles of the analysis are considered in finding the next most important variable.

**Availability.** Equiries on the package should be sent to SISSAD, Viale Campi Elisi 62, Trieste, Italy.

**Acknowledgements.** I thank the Hungarian Academy of Sciences, the University of Western Ontario, London, Canada, and the L. Eötvös University, Budapest, for the research facilities made available to me.

## REFERENCES

ANDERBERG, M.R. 1973. *Cluster Analysis for Applications*. Academic Press, New York.

BRUYNOOGHE, M. 1981. Classification ascendante hierarchique des grands ensembles de données: un algorithme rapide fonde sur la construction des voisinages reductibles. Les Cahiers de l'Analyse des Données 3: 7-33.

DAY, W.H.E. 1981. The complexity of computing metric distances between partitions. Math. Soc. Sci. 1: 269-287.

DIDAY, E., J. LEMAIRE, J. POUGET and F. TESTU. 1982. *Elements d'Analyse de Donnees*. Dunod, Paris.

DIGBY, P.G.N and R.A. KEMPTON. 1987. *Multivariate Analysis of Ecological Communities*. Chapman and Hall, London.

DOBSON, A.J. 1975. Comparing the shapes of trees. Lecture Notes in Mathematics 452: 95-100.

DOWNTON, M. and T. BRENNAN. 1980. Comparing classifications: An evaluation of several coefficients of partition agreement. Unpublished draft, Human Systems Institute, Boulder, Colorado. Abstract in: Classification Soc. Bull. 4 (4): 53-54.

DUBIEN, J.L. and W.D. WARDE. 1979. A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms. Canad. J. Stat. 7: 29-38.

FEOLI, E. and L. ORLOCI. 1979. Analysis of concentration and detection of underlying factors in structured tables. Vegetatio 40: 49-54.

FOWLKES, E.B. and C.L. MALLOWS. 1980. A method for comparing two hierarchical clusterings. Classification Soc. Bull. 4: 54.

GAUCH, H.G. 1977. *ORDIFLEX*. A flexible computer program for four ordination techniques: weighted averages, polar ordination, principal components analysis, and reciprocal averaging. Cornell University, Ithaca, N.Y.

GAUCH, H.G. and R.H. WHITTAKER. 1981. Hierarchical classification of community data. J. Ecol. 69: 537-557.

GOODMAN, L.A. and W.H. KRUSKAL. 1954. Measures of association for cross-classifications. J. Amer. Stat. Ass. 49: 732-764.

GORDON, A.D. 1980. On the assessment and comparison of classifications. In: R. Tomassone (ed.), *Analyse de Donnees et Informatique*. I.R.I.A., Le Chesnay, pp. 161-171.

GORDON, A.D. 1981. *Classification*. Chapman and Hall, London.

GORDON, A.D. 1987. A review of classification. J. Roy, Stat. Soc. A. 150: 119-137.

GOWER, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325-338.

GOWER, J.C. 1971. Statistical methods of comparing different multivariate analyses of the same data. In: F.R. Hodson, D.G. Kendall and P. Tautu, eds.), *Mathematics in the Archaeological and Historical Sciences*. Edinburgh Univ. Press, Edinburgh. pp. 138-149.

GOWER, J.C. 1975. Generalized Procrustes analysis. Psychometrika 40: 33-51.

GREENACRE, M.J. 1984. *Theory and Applications of Correspondence Analysis*. Academic Press, London.

JUHÁSZ-NAGY, P. 1984. Spatial dependence of plant populations. Part 2. A family of new models. Acta Bot. Hung. 30: 363-402.

JUHÁSZ-NAGY, 1985. A növények szerkezetvizsgálata: új modellek. IV. Problémafelvetés az autocönológiában. Bot. Közlem. 72: 1-15.

JUHÁSZ-NAGY, P. and J. PODANI. 1984. Information theory methods for the study of spatial processes and succession. Vegetatio 51: 129-140.

LANCE, G.N. and W.T. WILLIAMS. 1966. A generalized sorting strategy for computer classifications. Nature 212: 218.

LANCE, G.N. and W.T. WILLIAMS. 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. Computer J. 9: 373-380.

LANCE, G.N. and W.T. WILLIAMS. 1968. Note on a new information-statistic classificatory program. Computer J. 11: 195.

LEGENDRE, L. and P. LEGENDRE. 1983. *Numerical Ecology*. Elsevier, Amsterdam.

MARDIA, K.V., J.T. KENT and J.M. BIBBY. 1979. *Multivariate Analysis*. Academic Press, London.

ORLOCI, L. 1969. Information theory models for hierarchic and non-hierarchic classifications. In: A.J. Cole (ed.), *Numerical Taxonomy*, Academic Press, London. pp. 148-164.

ORLOCI, L. 1973. Ranking characters by a dispersion criterion. Nature 244: 371-373.

ORLOCI, L. 1976. Ranking characters by an information criterion. J. Ecol. 64: 417-419.

ORLOCI, L. 1978. *Multivariate Analysis in Vegetation Research*. 2nd ed. Junk, The Hague.

PIELOU, E.C. 1984. *The Interpretation of Ecological Data*. Wiley, New York.

PODANI, J. 1979a. Association analysis based on the use of mutual information. Acta Bot. Hung. 25: 125-130.

PODANI, J. 1979b. Generalized strategy for homogeneity-optimizing classificatory methods. In: L. Orloci, C.R. Rao and W.M. Stiteler (eds), *Multivariate Methods in Ecological Work*. Internat. Cooperative Publ. House, Fairland, Maryland, USA. pp. 203-209.

PODANI, J. 1980. SYN-TAX: Computer program package for ecological, coenological and taxonomical classifications. Abstracta Botanica 6: 1-158.

PODANI, J. 1982. Spatial processes in the analysis of vegetation. Ph. D. thesis. University of Western Ontario, London, Canada. Natl. Libr. of Canada Microfiche No. TC 56124.

PODANI, J. 1984a. Spatial processes in the analysis of vegetation. Theory and review. Acta Bot. Hung. 30: 75-118.

PODANI, J. 1984b. Analysis of mapped and simulated vegetation patterns by means of computerized sampling techniques. Acta Bot. Hung. 30: 403-425.

PODANI, J. 1984c. SYN-TAX II. Computer programs for data analysis in ecology and systematics. Abstracta Botanica 8: 73-94.

PODANI, J. 1986. Comparison of partitions in vegetation studies. Abstracta Botanica 10: 235-290.

PODANI, J. 1987. Computerized sampling in vegetation studies.

Coenoses 2: 9-18.

PODANI, J. 1988a. A method for generating consensus partitions and its application to community classification. Coenoses (in press).

PODANI, J. 1988b. New combinatorial SAHN clustering methods. (in press).

PODANI, J. 1984c. SYN-TAX III. User's manual. Abstracta Botanica 12. suppl. 1: 1-183.

PODANI, J. and T.A. DICKINSON. 1984. Comparison of dendrograms: a multivariate approach. Can. J. Bot. 62: 2765-2778.

RAND, W.M. 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Stat. Ass. 66: 846-850.

ROBINSON, D.F. and L.R. FOULDS. 1979. Comparison of weighted labelled trees. Lecture Notes in Mathematics 748: 119-126.

SCHOENEMANN, P.H. and R.M. CARROLL. 1970. Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika 35: 245-256.

SIBSON, R. 1978. Studies in the robustness of multidimensional scaling: Procrustes statistics. J. Roy. Stat. Soc. B. 40: 234-248.

SNEATH, P.H.A. and R.R. SOKAL. 1973. *Numerical Taxonomy*. Freeman, San Francisco.

WARD, J.H. 1963. Hierarchical grouping to optimize an objective function. J. Amer. Stat. Ass. 58: 236-244.

WILLIAMS, W.T., J.M. LAMBERT and G.N. LANCE. 1966. Multivariate methods in plant ecology. V. Similarity analyses and information analysis. J. Ecol. 54: 427-445.

WISHART, D. 1969. An algorithm for hierarchical classifications. Biometrics 25: 165-170.

WISHART, D. 1975. *CLUSTAN IC user's manual*. University College, London.