COENOSES 12(2-3):149-152, 1997
© C.E.T.A., Gorizia, Italy

Software abstract

SYN-TAX 5.1: A NEW VERSION FOR PC AND MACINTOSH COMPUTERS

J. Podani

Department of Plant Taxonomy and Ecology, Eötvös University, Ludovika tér 2, H-1083 Budapest, Hungary.
Email: podani@ludens.elte.hu, Web: http://ramet.elte.hu/~podani

SYN-TAX 5.1 is an improved version of the previous release, which was numbered **5.02** for both the IBM PC and compatibles and for Mac OS (Podani 1993, 1995). It offers many new methods, some of which not appearing in commercial program packages, whereas others greatly improving the educational usefulness of the package.

A Summary of New Features

Ordinal (nonmetric) clustering procedures, both agglomerative-hierarchical and iterative-nonhierarchical are provided. Three measures of resemblance for the ordinal data type are now available. The ordinal resemblance measures consider only the rank order of the data values and, accordingly, the clustering procedures also consider only the rank order of resemblances in the detection of group structure in the data. They are recommended for analyzing data with Braun-Blanquet and derived scale types commonly used in phytosociology, for example.

Constrained ordination methods, namely *canonical correspondence analysis* and *redundancy analysis* are included. The program provides eigenanalysis solutions as described in Jongman et al. (1987), and has therefore certain problem size limitations as compared to the iterative strategy used in **CANOCO** (ter Braak 1990). Graphics output includes triplots (joint ordination of observations, criterion and constraining variables).

Three types of *biplots* can be selected for display after principal components analysis. These are: Euclidean, Mahalanobis and mixed (Rohlf's) biplot.

A cladistic method, *neighbor joining* of Saitou and Nei, is added to produce optimally additive trees based on dissimilarity matrices. The additive tree is displayed graphically, just as ordinary dendrograms.

For the evaluation of digitized multi-species point patterns a new information-theory based method, *individual centered analysis* is implemented. In the DOS version all pattern analysis modules are now avaiable through a new shell program, **SYN-PATT**.

New *utility* programs – for the DOS version only – include routines to convert **NT-SYS** dendrogram files into **SYN-TAX** format and vice versa, and to convert NEXUS tree files into **SYN-TAX** format.

How to Upgrade?

DOS version. Upgrading **SYN-TAX 5.02** to ver. **5.1** is relatively easy. Copy the SYNUP51.EXE file into the SYNTAX directory, and then execute this program. It will decompress the new and modified .EXE and .OVR files, automatically replacing the old ones. Then, upon executing SYNTAX or SYNPATT the requested pictorial menu appears on the screen, and the program is ready to accept instructions.

Owners of even earlier versions of **SYN-TAX** should completely erase the SYNTAX directory, and perform a full installation from the distribution disks.

Macintosh version. The six old applications should be replaced by the new version completely. Applications **HIERCLUS, NONHIER, MULPATT** and **ORDIN** have been rewritten with several, self-explanatory changes in the contents of some menus, whereas only cosmetic changes were made in **EVAL** and **MATRANK.**

Theory in Brief

Ordinal Clustering

Resemblance coefficients offered by the previous versions can accept nominal, interval or ratio scale data; ordinal data were not allowed without transformation to the previous types. Now three coefficients for ordinal data are available, these are Kendall's tie-adjusted τ, Goodman and Kruskal's γ and a new measure of discordance recommended when presence/absence is also meaningful (e.g. phytosociological data with Braun-Blanquet scores). All of them are described in detail by Podani (1997b, this issue). The τ and γ functions are provided as complements, i.e. in dissimilarity form. These coefficients are available when ordinal non-hierarchical or ordinal hierarchical clustering is selected as the data analysis technique. If you wish to use them with non-metric multidimensional scaling or with single link or complete link clustering (with other clustering methods their use is not recommended), then you may save the dissimilarity matrix in ordinal hierarchical clustering and start the other analysis from this matrix, as usual in **SYN-TAX**. Since the computation of these coefficients is slow, it is always useful to save the dissimilarity matrix. There is another save option (DOS version only): the *ranks of the dissimilarity values* may be saved as a separate matrix. You may want to use this option if you wish to perform non-hierarchical ordinal clustering as well, since the time-consuming ranking process can be skipped in the second analysis.

The algorithm of ordinal clustering, both agglomerative and divisive, first orders the $m(m-1)/2$ coefficients, so that each d_{jk} is replaced by its rank, r_{jk} (Podani, in preparation). Then, the same clustering criterion is considered in both cases:

$$C = (R_w - R_{min}) / (R_{max} - R_{min})$$

where R_w is the sum of ranks of within-cluster dissimilarities, R_{min} is the possible minimum of such ranks for the given number of clusters and for the given numbers of objects in each cluster, and R_{max} is the possible maximum. A similar criterion – using R_{exp} instead of R_{max} – was proposed by the author (Podani 1997a) for determining the importance of variables in classifications; this method is already available in **SYN-TAX**. The value of C ranges from 0 to 1, 0 indicating that all within-cluster dissimilarities are smaller than the between-cluster dissimilarities, whereas 1 indicating that all between-cluster dissimilarities are smaller than the others.

In non-hierarchical clustering $1-C$ is maximized (i.e., C minimized) in each iteration step until no improvement is possible, whereas in agglomerative clustering C is minimized for each fusion. The result of agglomerative clustering for a given number of clusters, k, can usually be improved by non-hierarchical clustering with the same value of k.

In the dendrogram resulted from agglomerative ordinal clustering the ranks of fusions (values from 1 to m-1) are used, rather than the C values themselves, because this criterion does not change monotonically. That is, the result is an *ordered dendrogram* (Lapointe & Legendre 1995), rather than a weighted dendrogram, being consistent with the ordinality of the previous steps of the analysis.

Constrained Ordination

If you have two logically separable groups of variables in the data set, you may find useful two new options in **SYN-TAX** in addition to the *Canonical correlation analysis* option (CCA). Whereas in CCA the axes are obtained such that the two groups of variables are mutually constrained by each other, in *Redundacy analysis* (RDA) and *Canonical correspondence analysis* (CCOA) the ordination axes for one set of variables, the criterion variables, are constrained by the other set, but not vice versa. A common example is ecological: one domain is the species set, whereas the constraining variables are environmental descriptors. In the ordination of objects for the species data the axes must be linear combinations of the environmental variables. Redundancy analysis is the constrained form of standardized principal components analysis, whereas CCOA is a constrained form of correspondence analysis (COA). The first method is recommended for relatively linear data structures (such as those often observed along short gradients) and the second is said to be more suitable to long gradients. ter Braak & Prentice (1988) give an excellent summary of the theory. The program is based on eigenanalysis, as described in Jongman et al. (1987), another useful text. ter Braak & Verdonschot (1995) give many hints on the interpretation of CCOA results. Birks et al. (1996) present a full bibliography of the applications up to 1993.

Both methods are available through the **Ordination/ Canonical Analysis** menu (DOS) or the **Analysis** menu item (Macintosh). Be careful that the data are arranged such that variables are columns and observations (sites, objects) are in rows. The criterion variables must be given first, followed by the constraining variables in the subsequent columns. The number of the latter can never exceed the number of the former, a condition easily satisfied. In ecological ordinations, for example, we usually have much more species than environmental variables. For CCOA there are the same three weighting options as for COA. In most cases site scores are calculated such that they are in the barycenter of species scores (i.e., sites scores are weighted averages of species scores), although the other two options may also be useful (see references cited above).

The graphic result of RDA and CCOA is a scattergram showing two kinds of variables and the objects simultaneously, hence it is called the *triplot* (Šmilauer 1990). The two types of variables appear in different colours to facilitate interpretation. In the RDA result arrows point to all variables, whereas in the CCOA diagrams the arrows point to the environmental variables only. Labels can be used to identify the points just like in other **SYN-TAX** ordination scattergrams.

PCA Biplots

In the new version of the program the user has a choice among three types of PCA biplots if covariance or correlation is used (centered and standardized PCA, respectively):

- distances between points approximate the Euclidean distances among objects, whereas variable scores are the eigenvectors;
- interpoint distances approximate Mahalanobis generalized distances of objects, variable scores are covariances or correlations;
- interpoint distances approximate the Euclidean distances of objects, whereas variable scores are covariances or correlations (mixed or Rohlf biplot).

For more information on these biplot types, see e.g., Marcus (1993). In the previous version of **SYN-TAX** only the last option was available such that correlation was shown for variables in case of both centered and standardized PCA. In the new version this ambiguity is solved and there are more choices for the convenience of the user.

Neighbor Joining

This is a cladistic method proposed by Saitou & Nei (1987) to generate optimally additive trees from distance/

dissimilarity matrices. The steps of the analysis are described, for example, in Swofford & Olsen (1990). In the resulting tree the sum of lengths of branches along the path between any two objects approximates their input distance. The analysis produces an unrooted tree, which may be displayed graphically like minimum spanning trees in **SYN-TAX**. However, the program may also be instructed to yield a rooted tree, with two options to define the position of the root:

- *Outgroup rooting.* To achieve this, you must add an extra object, the outgroup object, to the data set. In taxonomic analysis, for example, the outgroup object is an external taxon which is the closest to the group being analyzed. The root will be positioned on the branch that connects this outgroup object to the remaining objects.
- *Midpoint rooting.* The two objects that are furthest apart are identified and then the root is positioned halfway between them (Figure 1).

There is no built-in routine to compute the distances from raw data. The distance matrix needs to be computed and saved previously via hierarchical clustering, such as single linkage, or in principal coordinates analysis. The tree diagram is automatically displayed after the analysis. The saved tree file can be used to reproduce the diagram at later time.

Individual-centered Analysis of Mapped Point Patterns

This method, proposed recently by Podani & Czárán (1997) is a spatial analysis of mapped and digitized point patterns representing multi-species assemblages. The procedure is based on information theory, and in some sense can be considered as the multivariate generalization of Ripley's *count distance method*. The analysis draws circles around each individual with increasing radii and examines whether the species combinations found in these sample units are commoner or rarer than expected given the assumption of complete spatial randomness. The statistics drawn in the function of radius, either for the whole set of species or for separate species, are informative as to the spatial "behaviour" of the point pattern analyzed.

The *Macintosh version,* as implemented in the application **MULPATT**, automatically draws the diagrams.

For the *DOS version* – based on the saved results – the user may draw the plots using any commercial graphic package. (The DOS program's name is **DARIUS**, after the rich Persian king – referring to species "richness" –, which is at the same time the permutation of the word "radius".) Individual centered analysis and all other simulated sampling and pattern analysis procedures are no longer available from the main shell of **SYN-TAX**. A new shell program, **SYN-PATT** was created to serve this purpose. In this way, the multivariate data analysis procedures have been separated from the multivariate pattern analysis procedures. In the main menu of **SYN-TAX** the pattern analysis option is still shown, but after choosing it the user receives a warning that the other program must be started. However, **SYN-TAX** and **SYN-PATT** are recommended to be kept together in the same directory, preferably named SYNTAX.

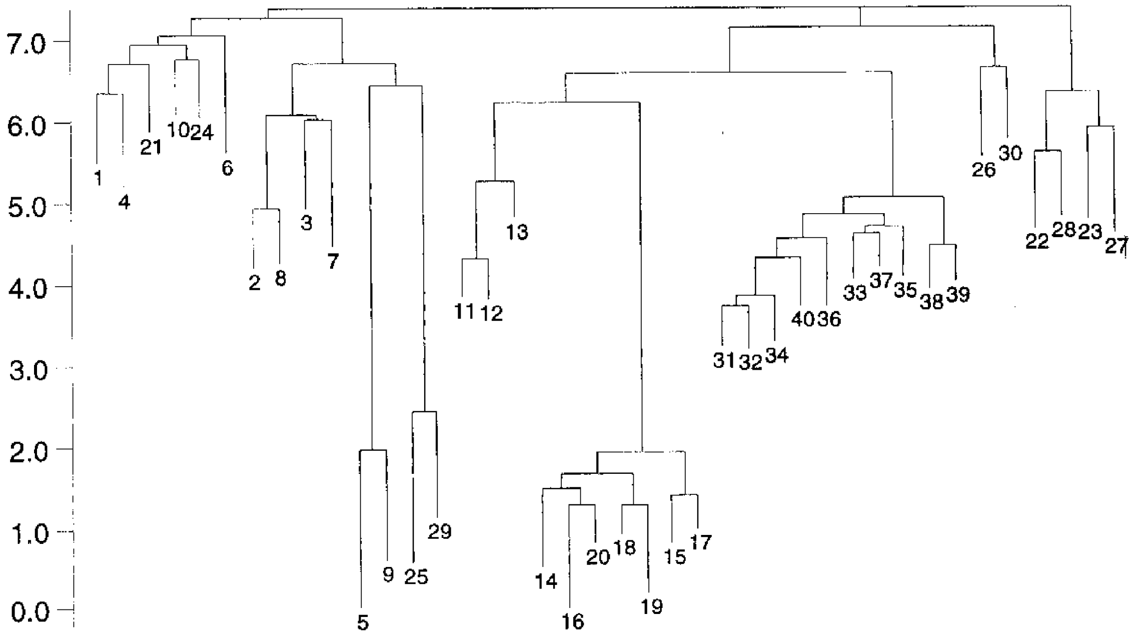

Figure 1. Graphics output example illustrating a midpoint-rooted neighbor joining tree as displayed by application HIERCLUS in SYN-TAX 5.1-Mac.

Tree Conversion Routines

SYN-TAX saves dendrograms in the format described in the User's Manual (Podani 1993). This format is not compatible with other tree formats, such as NEXUS commonly used in cladistic data analysis programs and the dendrogram format of the **NT-SYS** package (Rohlf 1993). Two new utilities – available only in the DOS version – solve this problem.

The program for converting NEXUS rooted tree files into a format readable by SYNTAX routines for tree plotting and comparison is NTOS.EXE. The input is a NEXUS file (as prepared by, say, **PHYLIP**) in which objects are labeled by *numbers* from 1 to *m* rather than by their names. Example:

```
((1:0.03947,(2:0.00775,(3:0.04082,
4:0.03416):0.01714):0.03238):.0411,5:0.02883);
```

In other words, a file specified as shown below cannot be used:

```
((Orangutan:0.03947,(Man:0.00775,(Gorilla:0.04082,
Chimpanzee:0.03416):0.01714):0.03238).0411,Gibbon:0.
02883);
```

unless labels are replaced by numbers.

The output is a SYNTAX file with *m*-1 rows and 5 columns. The fifth column contains a dummy hierarchical level which is simply the number of objects in the cluster. The result for the above example is:

```
3 4 1 1 2
2 3 1 2 3
1 2 1 3 4
1 5 4 1 5
```

Please note that subsequent comparisons of trees may use only the following descriptors of Podani (1982) and Podani & Dickinson (1984): topological difference, subtree membership divergence and cluster membership divergence, or any combination of these. Do NOT use cophenetic difference and partition membership divergence as they are meaningless because of the arbitrary levels. That is, we do not utilize edge lengths of the input tree.

Program **DENTRANS** has been written for for converting dendrogram data as output by program **NT-SYS** into **SYN-TAX** format and vice versa.

References

Birks, H. J. B., S. M. Peglar & H. A. Austin. 1996. An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986-1993. Abstracta Botanica 20:17-36.

Jongman, R. H. G., C. J. F. ter Braak & O. F. R. van Tongeren (eds.), 1987. Data Analysis in Community and Landscape Ecology. Pudoc, Wageningen.

Lapointe, F.-J. & P. Legendre. 1995. Comparison tests for dendrograms: a comparative evaluation. J. Classification 12:265-282.

Marcus, L. F. 1993. Some aspects of multivariate statistics for morphometrics. In: Marcus, L. F. et al. (eds.), Contributions to Morphometrics, Monografias Museo Nacional de Ciencias Naturales, Madrid, pp. 95-130.

Podani, J. 1982. Spatial processes in the analysis of vegetation. PhD thesis, UWO, London, Ontario, Canada.

Podani J. 1993. SYN-TAX. User's Guide. Scientia, Budapest.

Podani, J. 1995. SYN-TAX version 5.02 Mac. User's Guide. Scientia, Budapest.

Podani, J. 1997a. Explanatory variables in classifications and the detection of the optimum number of clusters. In: Hayashi, C. et al. (eds), Data Science, Classification and Related Methods. Springer, Tokyo.

Podani J. 1997b. A measure of discordance for partially ranked data when presence/absence is also meaningful.Coenoses 12:127-130.

Podani, J. & Czárán, T. 1997. Individual-centered analysis of mapped point patterns representing multi-species assemblages. J. Veg. Sci. 8:259-270.

Podani, J. & T. A. Dickinson. 1984. Comparison of dendrograms: a multivariate approach. Can. J. Bot. 62:2765-2778.

Rohlf, F. J. 1993. NTSYS-pc. Numerical Taxonomy and Multivariate Analysis System. Exeter Software, Setauket, NY.

Saitou, N. & M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

Šmilauer, P. 1990. Program CANODRAW, version 2.10. Scientia, Budapest.

Swofford, D. L. & G. J. Olsen. 1990. Phylogeny reconstruction. In: D. M. Hillis & C. Moritz (eds), Molecular Systematics. Sinauer, Sunderland, pp. 411-501.

ter Braak, C. J. F. 1990. Update notes: CANOCO version 3.10, Wageningen.

ter Braak, C. J. F. & I. C. Prentice. 1988. A theory of gradient analysis. Adv. Ecol. Res. 18:271-317.

ter Braak, C. J. F. & P. F. M. Verdonschot. 1995. Canonical correspondence analysis and related multivariate techniques in aquatic ecology. Aquatic Sciences 57/3.