SYN-TAX 2000

Computer program for data analysis in ecology and systematics

for WINDOWS 95, 98 & NT

User's Manual

J. Podani, 2001

First print in April, 2001

ISBN 963 8326 23 9
Scientia Publishing, P.O. Box 658, H-1365 Budapest, Hungary

Contents

Chapter 1

Introduction

This is the most recent version of the SYN-TAX program package designed for multivariate data analysis in *SYN*biology (or ecology) and *TAX*onomy (or systematics), noting that the routines equally apply to problems outside in biology where many objects have been described in terms of many variables and the objective is exploration of inherent data structure.

1.1 Main features

System requirements: the package runs under WINDOWS 95, 98 and NT.

Modular structure: five independently functioning routines. The first release contains **HierClus, Ordin** and **NonHier.**

Module **HierClus**: Hierarchical clustering (h-SAHN and d-SAHN methods, information theory clustering based on entropy and mutual information, divisive monothetic clustering, ordinal clustering, neighbor joining, global optimization, minimum spanning trees). Use of 33 dissimilarity coefficients, 6 run-time standardization procedures. Dendrograms, minimum spanning trees, additive trees, unrooted trees. Utilities.

Module **Ordin**: Metric and nonmetric ordination including PCA, PCoA, CA, CCoa, RDA, NMDS, canonical correlation analysis and canonical variates (discriminant) analysis. Ordination scatterplots, biplots, triplots, rotating plots. Minimum spanning trees and classification polygons superimposed on ordinations. Scree plots and Shepard diagrams. Use of 33 dissimilarity coefficients, 6 run-time standardization procedures in PCoA and NMDS. Utilities.

Module **NonHier:** Non-hierarchical clustering methods including *k*-means, multiple partitioning, global optimization, quick clustering, fuzzy c-means clustering and ordinal methods. Ternary plots and scattergrams to illustrate fuzzy classifications graphically. Use of 33 dissimilarity coefficients, 6 run-time standardization procedures in global optimization, three rank coefficients for ordinal clustering. Utilities.

Module **Eval**: not yet available

Module **MatRank:** not yet available

1.2 Major changes compared to ver. 5.1

- Switch from DOS to 32-byte WINDOWS systems.
- No program-dependent problem size limits.
- Input file header now includes title and matrix size information.
- New format for label files. Long labels allowed in ordination scattergrams.
- More flexible graphics options, called from the main menu or in context-dependent form as pop-up menus when a diagram is displayed.Improved graphics features.
- Zooming entire diagrams or portions thereof to examine small details.
- New graphics formats: BMP, EMF and WMF.
- TIF and PCX no longer supported, except in rotating plots.
- Export/import of EXCEL datafiles.
- Simple text editor/file viewer.
- Only interactive mode; batch no longer available.

- Data grid in main window, showing input matrix.
- Analysis of point patterns not included.
- New utilities to create bootstrapped, randomized or otherwise resampled data files.
- Analysis of a data matrix in two ways, without previously transposing.
- Minimum spanning and unrooted neighbor joining trees with branch lengths.
- Scree plots for eigenvalues.

1.3 History

SYN-TAX 2000 is a result of a continuous programming project running since 1980. The first version was made for mainframe computers and performed only hierarchical clustering [1]. The package was considerably expanded to include ordination methods and then various procedures for comparing and evaluating clustering and ordination results [2]. A main step forward is the appearance of the first version for a personal computer [3-9], providing access to all previously programmed routines under MS-DOS. SYN-TAX IV was still a command-based DOS program [10] subsequently incorporated into a graphical shell in version 5.0 [11-13]. The last upgrade to SYN-TAX is version 5.1, containing neighbor joining, canonical correspondence analysis and ordinal clustering [14].

1.4 Selected references on SYN-TAX versions

[1] Podani, J. 1980.Computer programs for ecological and taxonomical classifications (in Hungarian). *Abstracta Botanica* 6:1-158.

[2] Podani, J. 1984. SYN-TAX II. Computer programs for data analysis in ecology and systematics. *Abstracta Botanica* 8:73-94.

[3] Podani, J. 1988. SYN-TAX III. A package of programs for data analysis in community ecology and systematics. *Coenoses* 3:111-119.

[4] Podani, J. 1988. Graphics routines for SYN-TAX III. *Abstracta Botanica* 12:183-188.

[5] Podani, J. 1988. SYN-TAX III. User's Manual. *Abstracta Botanica* 12. Suppl. 1:1-183.

[6] Podani, J. 1989. SYN-TAX. Computer programs for data analysis in ecology and systematics. *Journal of Classification* 6:273-278..

[7] Podani, J. 1990. SYN-TAX III-pc - Supplement 1: Minimum spanning trees. *Abstracta Botanica* 14:1-6.

[8] Podani, J. 1990. SYN-TAX III-pc - Supplement 2: Fuzzy clustering. *Abstracta Botanica* 14:7-22.

[9] Podani, J. 1990. SYN-TAX III-pc - Supplement 3: Macintosh version. *Abstracta Botanica* 14:23-29.

[10] Podani, J. 1991. SYN-TAX IV. Computer programs for data analysis in ecology and systematics. In: E. Feoli & L. Orlóci (eds.), *Computer Assisted Vegetation Analysis*. pp. 437-452. Kluwer, The Netherlands.

[11] Podani, J. 1993. SYN-TAX 5.0: Computer programs for multivariate data analysis in ecology and systematics. *Abstracta Botanica* 17:289-302.

[12] Podani, J. 1994. *Multivariate Data Analysis in Ecology and Systematics*. SPB Publishing, The Hague. pp. 316.

[13] Podani, J. 1995. SYN-TAX 5.0: Computer programs for multivariate analysis in ecology and systematics. In: G. Guariso & A. Rizzoli (eds.), *Software per l'Ambiente*. Patron, Bologna. pp. 37-43.

[14] Podani, J. 1997. SYN-TAX 5.1: A new version for PC and Macintosh computers. *Coenoses* 12:149-152.

1.5 Further reading

Although some brief information on background theory is provided in this manual, the use of SYN-TAX 2000 requires a deeper knowledge of theoretical details. In addition to well-known texts in the subject area of clustering and ordination, two books by the author, one in Hungarian and the other in English, may be recommended:

Podani J. 1977. *Bevezetés a többváltozós biológiai adatelemzés rejtelmeibe*. Scientia, Budapest.
http:\\www.ramet.elte.hu\~scientia

Podani, J. 2000. *Introduction to the Exploration of Multivariate Biological Data*. Backhuys, Leiden.
http:\\ www.backhuys.com

1.6 Acknowledgements

I thank Zs. Erdei for his assistance in developing SYN-TAX 2000 and for writing the EXCEL conversion routine. J. Schumacher is thanked for encouragement and help. I would like to express my sincerest thanks to D. Walker, Dr. O. and M. Vitousek, Z. Naar and to all the participants of a workshop held in Taipei, Taiwan, February 2001, for their useful comments, patience and tolerance during beta testing of SYN-TAX 2000.

1.7 Correspondence

The author welcomes any questions, comments and suggestions for future improvement of the package at the following email address:

podani@ludens.elte.hu

Answers to frequently asked questions, upgrade notes and other relevant information will be available at the web site

http:\\ramet.elte.hu\~podani

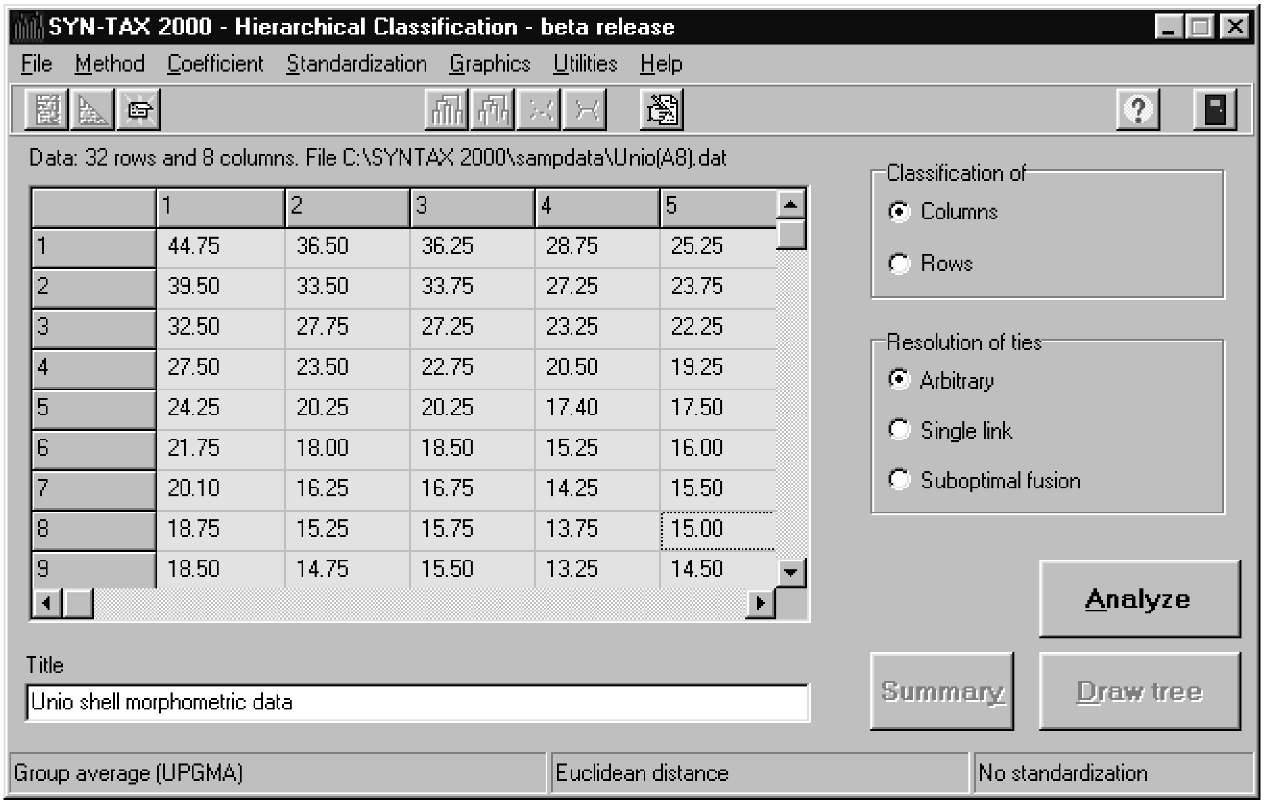
Reprints reporting on the use of SYN-TAX are greatly appreciated. See the postal address in the **About** box.

Chapter 2

Quick start

2.1 The Main Window

The Main Window of each module of SYN-TAX 2000 provides the **Main Menu** bar, the **Speed Button**s, the **Data Grid** with filename information, the **Title Box**, the **Status Bar**, some groups of **Radio Buttons** and the **Main Buttons**, as exemplified below by the Main Window of **HierClus.**



Data: 32 rows and 8 columns. File C:\SYNTAX 2000\sampdata\Unio(A8).dat

	1	2	3	4	5
1	44.75	36.50	36.25	28.75	25.25
2	39.50	33.50	33.75	27.25	23.75
3	32.50	27.75	27.25	23.25	22.25
4	27.50	23.50	22.75	20.50	19.25
5	24.25	20.25	20.25	17.40	17.50
6	21.75	18.00	18.50	15.25	16.00
7	20.10	16.25	16.75	14.25	15.50
8	18.75	15.25	15.75	13.75	15.00
9	18.50	14.75	15.50	13.25	14.50

Classification of
● Columns
○ Rows

Resolution of ties
● Arbitrary
○ Single link
○ Suboptimal fusion

Title
Unio shell morphometric data

Group average (UPGMA) | Euclidean distance | No standardization

The **Menu Bar** includes the main menus with the most important options. When data files are open, it is advisable to make your choice in the order of **Method, Coefficient**, and **Standardization**.

The **Speed Button**s, located in the Toolbar right below the Menu Bar, are useful to perform quick operations. The meaning of each button is clarified by a hint appearing when you place the cursor over it. Speed Buttons that cannot be used in the given context of your options are disabled. For example, graphics speed buttons cannot be used when data files are open, while the **Close input file** speed button is inactive when no files are open.

The name of the currently open file is given below the Speed Buttons with a full path. Further down you find the **Data Grid**, a scrollable window containing the data values. The rows and columns are numbered or labeled, if labels are opened for either or both, using commands from the **File** menu.

When a data file is open, the title (i.e., the first row in the file) is reproduced in the **Title** Box.

The **Status Bar** on the bottom of the Main Window informs the user about the currently selected options for analysis. When the program is started or, occasionally, when a radically new combination of options is chosen, the Status Bar displays the default options automatically.

The **Main Button**s located on lower right are used to start the analysis (**Analyze**) or to display text output (**Summary**) and graphics (all others, such as **Draw tree, Draw biplot**, etc.).

Some groups of **Radio Button**s provide further options, for example, to decide whether rows or columns in the data matrix are chosen as objects of the analysis. These buttons are disabled or hidden when their use is not allowed.

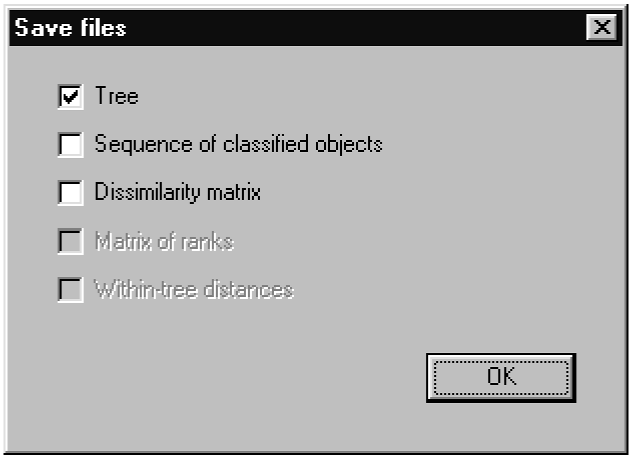
2.2 Data grid

After opening a data file, or a distance matrix file, the input values are displayed in a data grid within the main window. The background color may be changed from the **Utilities/Data Grid color** menu. If you do not want to see the data because, for example, the data file is too large thus causing memory problems, the data grid may be deactivated from the **Utilities/General options** dialog, by unchecking the **Data grid shown in the main window** box.

> **Note:** The contents of the cells of this grid cannot be edited, any modification of data should be carried out beforehand, either by Excel or using the built-in Text editor of SYN-TAX.

2.3 The File menu

The leftmost item in the menu bar of the main window is **File**. Commands of this menu can be used to open and close data and distance matrix files, to open and close labelfiles for rows and columns of datafiles, as will be exemplified in several places of this manual. (For data and distances, open and close operations are a step faster with speed buttons.) In addition, this menu includes the **Save output files** command which evokes a dialog box to specify what kind of partial results should be saved by the program. Such partial results include tree files, ordination score files, distance matrices and so on. This command is active only if there is a data or distance matrix open, and its items are enabled depending on your choice of the data analysis method. For example, **HierClus** has the following Save files dialog box:



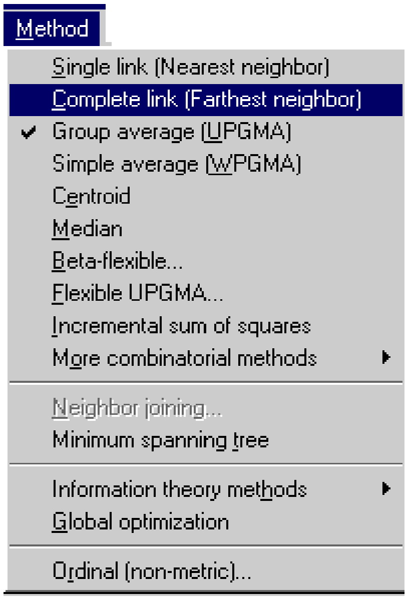
☑ Tree
□ Sequence of classified objects
□ Dissimilarity matrix
□ Matrix of ranks
□ Within-tree distances

The standard **Printer setup** and **Exit** commands conclude the **File** menu. The **Exit** command has its corresponding speed button at the rightmost position in the **Main Window**.

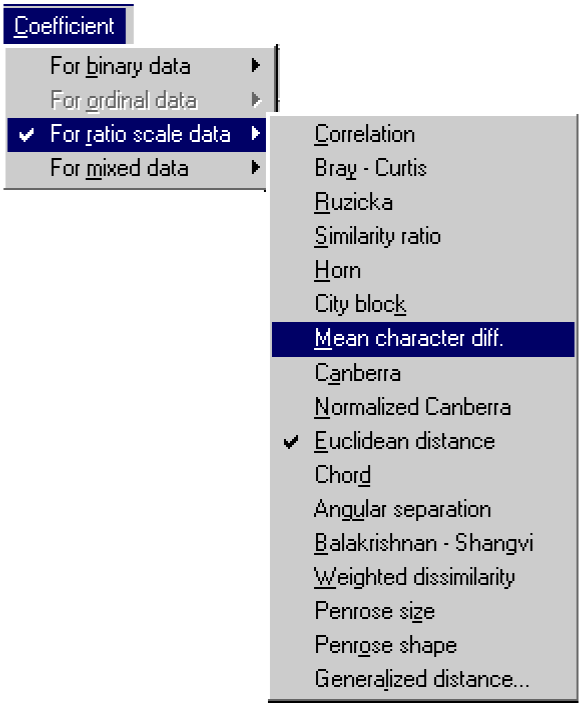
2.4 Your first session in SYN-TAX 2000

Double click the icon of **HierClus, NonHier** or **Ordin**. The Main Window appears. In this, most menu items are disabled, the program awaits your instructions. To continue, there are two possibilities:

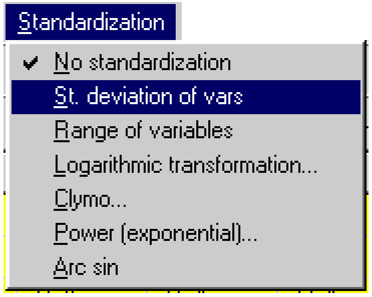
1) Open a data matrix and, possibly, label files, by commands from the **File** menu. Use the supplied sample data files first to see the main features of the program. The data appear in a grid within the main window. Then, the **Method, Coefficient** and **Standardization** menus become activated. Make your own choice of options *in that order*. For **HierClus**, this three-step procedure may be the following



(The contents of the **Method** menu depend on the module you are using.)



and



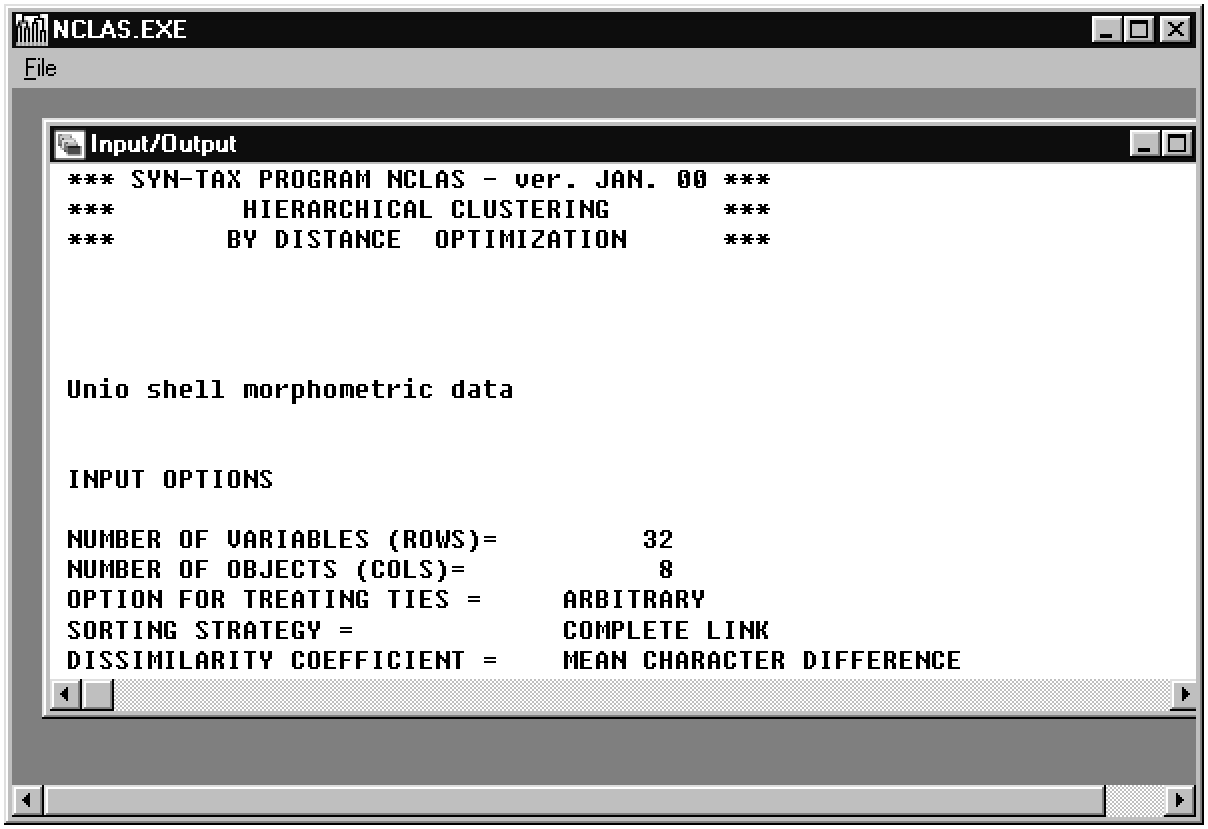
The options actually selected are always shown in the status bar on the bottom of the main window. In this case:

Complete link, Mean character difference, Standard deviation of vars.

should appear in the status bar. The respective menu items are checked for the convenience of the user. If satisfied with your choice, press the **Analyze** button.

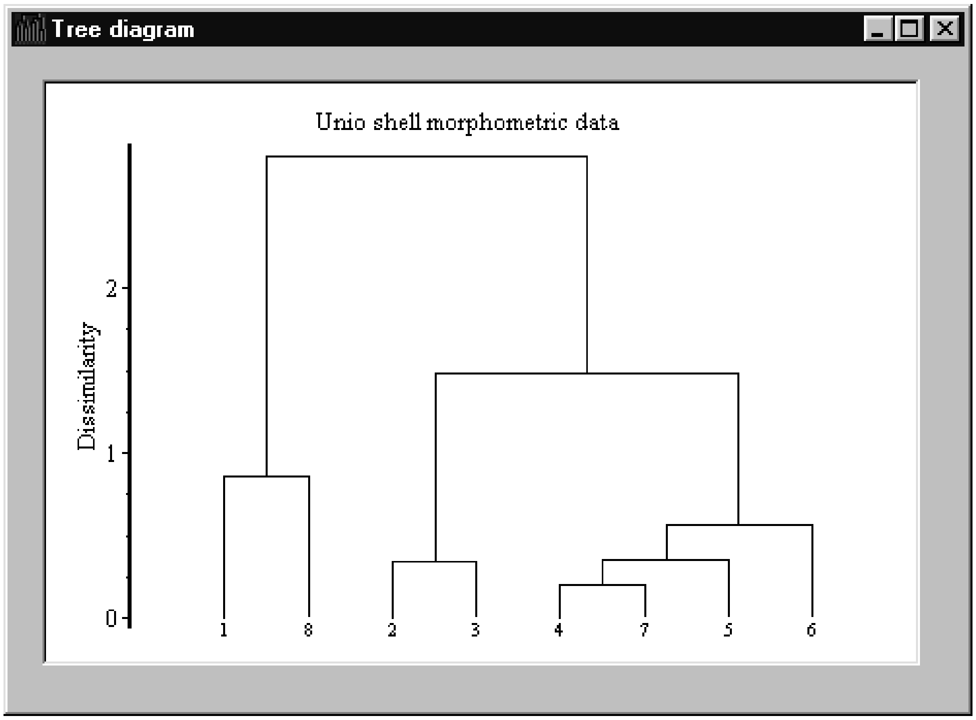
Depending on the options chosen, several new windows prompt you to specify input/output filenames and further details of the analysis.

Then, the computations start, with the numerical results appearing in a new output window. The output list contains all input parameters selected previously or accepted as default. You may save or print the contents, or close this window, after the output is typed:



```
*** SYN-TAX PROGRAM NCLAS - ver. JAN. 00 ***
***        HIERARCHICAL CLUSTERING       ***
***       BY DISTANCE  OPTIMIZATION      ***




Unio shell morphometric data


INPUT OPTIONS

NUMBER OF VARIABLES (ROWS)=          32
NUMBER OF OBJECTS (COLS)=             8
OPTION FOR TREATING TIES =     ARBITRARY
SORTING STRATEGY =             COMPLETE LINK
DISSIMILARITY COEFFICIENT =    MEAN CHARACTER DIFFERENCE
```

Upon clicking the activated button '**Draw Tree**' output found on lower right in the Main Window, a graphics window appears displaying the dendrogram, using a default window size.



2) If you already have a tree diagram or an ordination result saved in the appropriate SYN-TAX format (some examples are supplied together with the package, with self-explanatory names, Subsection 4.1.1) then you can directly press one of the graphics speed buttons or select the corresponding command from the **Graphics** menu. After opening a graphics file and (optionally) label file(s) (press **Cancel** if you do not want to use labels), the graphics is reproduced in a new window.

For example, let us reproduce the minimum spanning tree graphics from file MinSpTree20.mst. Click the speed button called **Draw minimum spanning tree from file**, or use the **Graphics/Draw Min sp. tree** command. Then, open the tree file and then cancel the **Open label file** window (so that the objects are only numbered in the diagram). The tree appears in the graphics window. Its properties may be changed by the commands available from the popup menu which appears if you click the right mouse button with the cursor placed over the drawing area.

Chapter 3

Input specifations, file conversions

SYN-TAX 2000 has many conventions for data input. In order to analyze the data correctly and to minimize the chance for input errors, the user must be familiar with the details that follow in this chapter. Note, first, that the last row of data files must be closed by a line feed character, achieved most conveniently by adding an empty line.

3.1 Raw data files

A data file contains a rectangular array of real or integer numbers, the raw data matrix to be analyzed. In general, SYN-TAX 2000 requires that data should be prepared in text files such that

- the data values are separated by at least one space or linefeed character,
- the first row contains a title,
- the second row contains information on matrix size,
- in the file, each new row of the data matrix should start in a new line,
- any row of the data matrix can be continued in as many lines as necessary.

Appearance of commas, instead of decimal points, and letters in the main body of data will abort the execution of the program. The **Analyze** button in the Main Window is disabled if no file containing data (or distances) is opened. Data matrices can be opened by the **Open raw data** command from the **File** menu or the speed button with the data matrix icon:

Note: Before importing data formerly processed by SYN-TAX 5.1 or earlier versions, the only change to be made is to insert the title line and the parameters of the second row by the Text Editor.

The default extensions of data files are .DAT and .DTA. Data files have several subtypes, depending on whether there is a logical grouping of variables or objects.

3.1.1 Ungrouped raw data

Most commonly, a file of raw data is input with *n* rows and *m* columns, without any *a priori* grouping of variables or objects. In this case,

- the second row contains two numbers, *n* and *m*,

An example data file with *n* = 5, *m* = 6 and the second row of the matrix broken into two lines is given below:

```
Sample data set
5 6
1 0 4 1  2.3  4
5  2  3
5 7 0
0 3 4 5 6 0.88
1 3 0 0 4 5
4 0 6 1.2 0 0
```

3.1.2 Data with two groups of variables

Various forms of canonical analysis (CCoA, RDA, COR) in **Ordin** require that the *variables* are categorized into two logically distinct groups. In these cases, variables must always be provided as columns, whereas observations as rows in the file (the **Ordination of columns/rows** radio button is inactive if CCoA, RDA or COR is chosen). The criterion variables (e.g., species) are in the left, the explanatory variables (e.g., environmental variables) appear in the right columns. The second row of the header of the data file therefore contains

- four values: no. of observations (rows), total no. of variables (columns), no. of left domain vars. and no. of right domain vars.

Example:

```
Only to show how to prepare data for RDA, CCoA and COR
5 7 4 3
1 2 3 2 3.2 2.5 3.9
0 1 0 1 1.9 3.4 4.6
1 2 1 2 5.6 7.8 7.8
2 1 0 1 1.4 6.7 8.9
0 2 1 1 3.4 5.6 7.8
```

Hint: If your datafile does not contain the third and fourth numbers in the header, i.e. it is a simple raw data matrix, and the file contains *real* numbers, then you may still start CCoA, RDA or COR. In this case, the program will prompt you with a dialog box to define the number of *left* domain variables.

3.1.3 Data with several groups of objects

Canonical variates analysis (CVA, CANOVAR or discriminant analysis) in **Ordin** requires a data matrix in which the *objects* are grouped *a priori*. In this case, variables must always be provided as columns, whereas observations (objects) as rows in the file (the **Ordination of columns/rows** radio button is now inactive). For CVA,

- the first three values in the second row of the data file are: no. of observations (rows), no. of variables (columns), and no. of object groups. Then, in the same line follow the group size values.

In the main body of the data file the objects must be grouped and arranged in the same order as the group size values! An example for 20 objects, 7 variables and 3 groups, with group sizes 9, 5 and 6, is as follows:

```
CVA sample data
20   7   3   9   5   6
1 2 3 2 3 4 3
1 2 3 2 1 2 3
1 2 2 3 2 1 2
1 2 2 3 1 1 3
1 2 3 4 2 2 3
1 2 2 2 1 1 2
1 2 3 1 2 3 1
1 2 3 4 5 4 3
2 1 2 2 2 2 3

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

1 2 3 4 4 5 6
3 4 3 4 5 5 5
```

```
1 2 2 2 4 5 6
1 1 4 4 4 4 6
1 2 3 4 2 5 5
9 9 9 5 5 5 5
```

For clarity, optional linefeed characters may be inserted between groups, as seen above.

> **Hint:** The optimal arrangement of data is as shown above. However, in certain situations you may have a simple raw data matrix without grouped arrangement, and you still want to try CANOVAR. In this case, open the raw data matrix, start the analysis, and then you will be prompted for a filename containing group memberships (see Section 3.3, below). In the file of group memberships the number of values must be the same as the number of objects in the data file. The ith value identifies the group membership of the ith object in the data. The data matrix is then rearranged by the program automatically. Caution: IN THE OUTPUT LIST, THE ID. NUMBERS OF THE OBJECTS ARE THOSE AFTER THE REARRANGEMENT, AND NOT THE ORIGINAL ONES!!!

3.2 Distance/dissimilarity matrices in input files

Certain procedures in SYN-TAX (clustering, multidimensional scaling) can analyze distance or dissimilarity matrices directly such that the original data are not needed. The standard format is the semimatrix form, with main diagonal included, as required by earlier versions of SYN-TAX. A substantial change compared to ver. 5.1 is that the file must contain a title in the first row and the number of objects (matrix size) in the second row. The number of rows in which the matrix is provided is immaterial, as in the earlier versions of the program.

The default extension of such files is .DIS.

An example for a 5x5 matrix:

```
Sample distance matrix
5
0
1 0
2 1 0
3 2 1 0
4 3 2 1 0
```

The **Analyze** button in the Main Window is disabled if no distance (or data) matrices are opened. Distance matrices can be opened by the **File/Open dis. matrix** command or the speed button with the semimatrix icon:

3.3 Group membership array (vector) defining a partition

Group membership arrays are single vectors of length *m*, where *m* is the number of objects classified in a partition. The *i*th value of the vector indicates the group to which object *i* belongs. In the file of group membership arrays, the first row contains *m*, potentially followed by a comment. Then follow the *m* group membership values starting in the next row and continued in an arbitrary number of rows. For example,

```
10    Any comment here
1 2 1 2 1 2 3 3 3 3
```

Group membership vectors are used in SYN-TAX 2000 to

- output results of non-hierarchical clustering;
- specify a partition of objects for Canonical Variates Analysis if, in the data file, no grouping of objects is supplied;
- specify a partition of objects in order to superimpose this classification using convex polygons on an ordination.

The default extension of partition files in SYN-TAX is .PAR.

3.4 Partition cluster seeds

For initializing non-hierarchical clustering, the user may wish to use seed objects for the starting clusters. These are provided in a small text file which has the default extension .PSD. An example for initializing an analysis of, say, 20 objects is

```
3
1  2  7
```

In this, 3 indicates that a 3-cluster partition will be derived, with objects 1, 2 and 7 as starting seeds. Of course, none of the id numbers can be larger than the number of objects to be analyzed.

3.5 Labels

Labels can be used to identify objects and variables in the typed output and graphics, as well as in the data grid of the main window. A label is a string of characters. The number of valid characters is limited to the first eight in most situations (dendrograms, minimum spanning trees, output lists), but in **Ordin** graphic displays the labels are output in as many characters as they appear in the file. The labels are to be provided in a separate file for the columns and in another file for the rows of the data matrix. Label files can be opened by the respective commands available in the **File** menu *before the analysis starts*. The format of the label file has GREATLY changed for this version. In this new file format, each label is provided in a separate line, and the first line of the file contains the number of labels, followed by and optional note. Example:

```
10 (A file with ten labels)
Savanna1
Savanna2
Desert1
Desert2
Desert3
Meadow1
Meadow2
Pasture1
Pasture2
Pasture3
```

Preferably, enter a blank line after the last label. This ensures that no input error appears when reading the labels.

The default extension of label files is .LAB in SYN-TAX 2000.

In order to use earlier versions of SYN-TAX label files, use the **Text editor** to insert the first line and to break the lines such that each new line has a single label. The program aborts if the number of labels is smaller than the number of data rows (or columns) for which a label file was opened earlier. Also, the program may abort if the last label is less than 8 characters and there is not an empty line after it.

> **Note**: If there are more than eight characters for a label to be used for tree diagrams, then the first eight characters are input only. Leading and trailing spaces are disregarded in displays.

To import labels from an Excel spreadsheet, just select the column block of the labels, then copy and paste the block into a text file. Then edit the text file, if necessary, using the built-in Text editor. This simple importing does not work if the labels are in a row of the EXCEL file. In this case, this row should be converted to a column beforehand.

> **Hint:** Data labels, as obvious from the above, are not supported to appear as row and column headers within data files. Too much confusion would arise otherwise (potential problems with input etc.). An advantage of providing separate label files is that the same diagram can be labeled in different ways: by a set of labels as text, and by another set of labels as symbols (in which case you can use Wingdings or other symbol fonts for the labels).

3.6 Converting file formats: distances

SYN-TAX can handle input distances/dissimilarities in semimatrix format only. If you have a full distance matrix from an external source, then its input to SYN-TAX may be solved by conversion. First, prepare the input file such that the first row contains a title, and the second contains matrix size. For example,

```
Sample full distance matrix
5
0 1 2 3 4
1 0 2 3 4
2 2 0 3 4
3 3 3 0 4
4 4 4 4 0
```

From the **Utilities** menu select the **Convert full dist. matrix** command. Open the full matrix and then specify a name for the output semimatrix which is then ready for analysis by SYN-TAX. The contents of the output file will be the following:

```
Semimatrix from Sample full distance matrix
5
0
1 0
2 2 0
3 3 3 0
4 4 4 4 0
```

There is an option for conversion in the opposite direction. To export a semimatrix to full format, open first a SYN-TAX dissimilarity matrix. Then, from the **Utilities** menu, select the **Export dist. matrix to full format** command, and follow the instructions. The output file will have the structure as shown above.

3.7 Transposing data

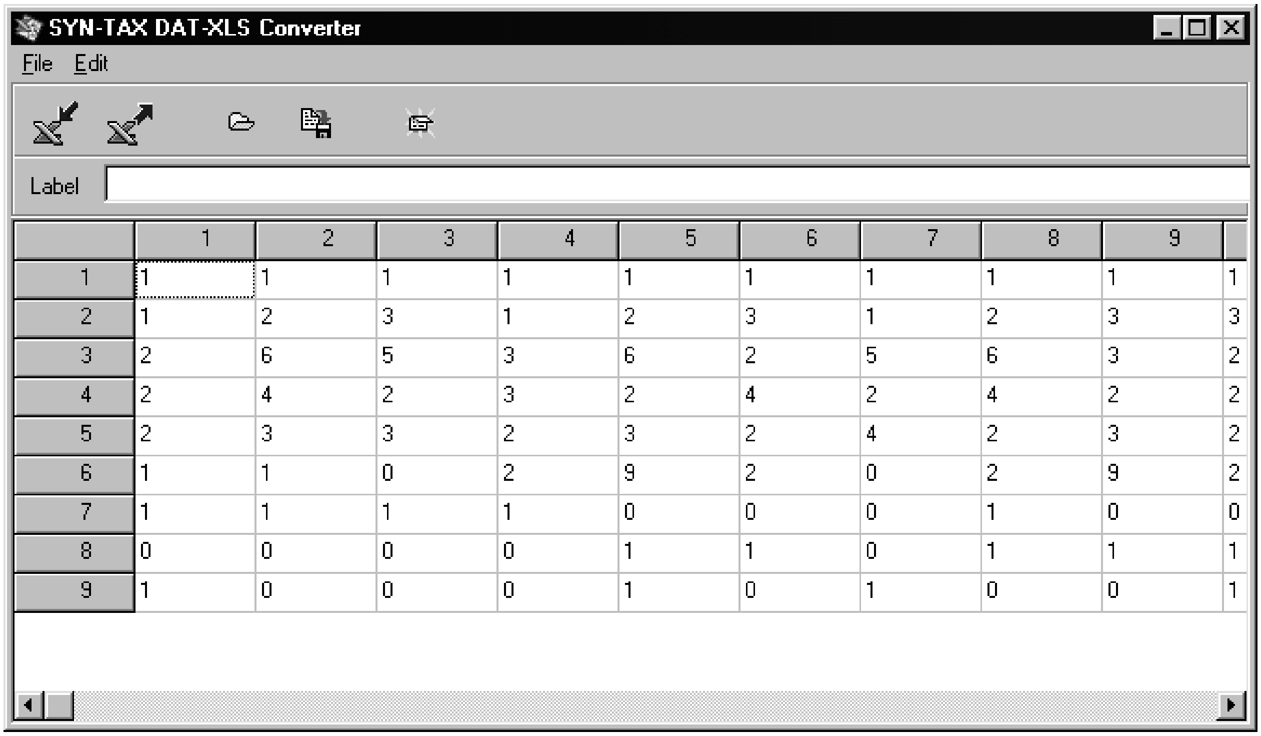
Although SYN-TAX has the flexible option for the analysis of either rows or columns of data arrays, a choice governed by radio buttons, you may want to transpose data matrices permanently (CCoA may require this, for example). In this case, use the **Utilities/Transpose data matrix** command. The output file prepared will then be ready for analysis by SYN-TAX 2000.

3.8 Excel files

Import/export data from/to Excel files is possible using the **Utilities/Excel import-export** command. This utility program can only be used to convert raw, *n* by *m* data matrices in both directions.

Before importing data from an Excel spreadsheet, please clear from the Excel table all rows and columns that contain only text or labels. Import can be successful if the ***main body of the data matrix only is present in the spreadsheet***. When the Excel file is prepared this way, then call the utility program, read the Excel file, and enter a title into the title box. Then save the data in DAT format. The converter will save the contents in the format required by SYN-TAX routines.

The Main Window of the converter is:



	1	2	3	4	5	6	7	8	9	
1	1	1	1	1	1	1	1	1	1	1
2	1	2	3	1	2	3	1	2	3	3
3	2	6	5	3	6	2	5	6	3	2
4	2	4	2	3	2	4	2	4	2	2
5	2	3	3	2	3	2	4	2	3	2
6	1	1	0	2	9	2	0	2	9	2
7	1	1	1	1	0	0	0	1	0	0
8	0	0	0	0	1	1	0	1	1	1
9	1	0	0	0	1	0	1	0	0	1

An alternative possibility is to **Copy** the requested block of values from the Excel spreadsheet, then open a file by the **Text Editor** and paste the data block into a new file.

Importing labels from Excel spreadsheets is possible by the copy and paste operation as described above. After that, editing the file using the **Utilities/Text Editor** built-in routine may be necessary to satisfy format requirements. Essentially, it means that the first line of the label file should contain the number of labels, optionally followed by a remark, and then from the next line follow the labels, one label per line.

Note: Use decimal points, rather than commas, in Excel as well.

Chapter 4

Graphics

A most essential part of any multivariate program package is its graphics interface. After analysis in SYN-TAX, the graphics results are displayed upon pressing graphics buttons. Alternatively, graphics displays may be reproduced from files saved in previous analyses. The following discussion provides details on file treatment, graphics formats and commands used in SYN-TAX 2000.

4.1 General features of SYN-TAX 2000 graphics

4.1.1 Graphics files

SYN-TAX saves graphics results in specific formats for future use by the program or for reproduction with various graphics options. Open the sample graphics files with the **Text Editor** and examine their contents for more information.

Sample graphics files supplied with the program include

Dendrogram20.DEN	(dendrogram for 20 objects)
RootedAddTree20.ADT	(rooted neighbor joining tree for 20 objects)
MinspTree20.MST	(minimum spanning tree for 20 objects)
UnrootedAddTree20.UTR	(unrooted additive tree for 20 objects)
ObjScores20.ORD	(PCA coordinates for 20 objects)
VarScores9.ORD	(PCA coordinates for 9 variables)
TriplotVars8.ORD	(RDA triplot variable coordinates)

4.1.2 Graphics formats, metafiles

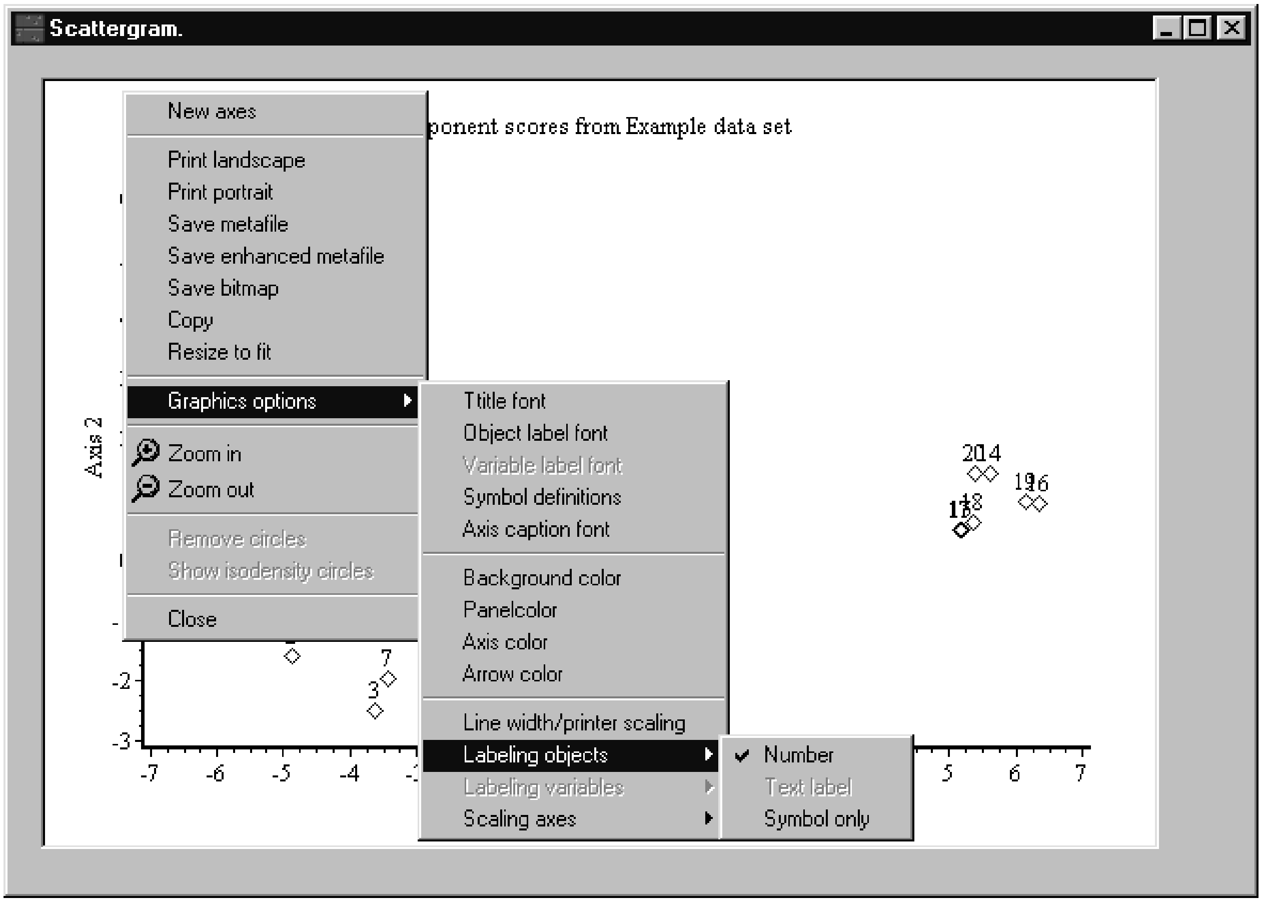
SYN-TAX 2000 supports three formats for graphics results. The **Copy** and **Save** commands available from the pop-up menus when a graphics window is active, can be used to copy and save the drawing in bitmap (BMP) format.

Ordination scattergrams, trees, scree plots and Shepard diagrams can be saved in Windows metafile format as well. To achieve this, press the **Save metafile** button (in case of scree plots and Shepard diagrams) or use the pop-up menu command **Save metafile**. This format ensures high-quality reproduction of the graphics result when printed later or embedded into a document. The diagrams saved in metafile (*.WMF) or enhanced metafile (*.EMF) format are editable item by item using an appropriate graphics program, such as Adobe Illustrator.

> **Note:** The TIF and PCX formats formerly available in SYN-TAX 5.1 and earlier versions are no longer supported, except in the rotating plot routine (Section 4.2).

4.1.3 Pop-up menus

When a graphics window is open (with trees, ordinations, etc.) a pop-up menu is evoked by clicking the *right* mouse button provided that the cursor is placed within the graphics area. The menu has a wide variety of commands (Section 4.3) for redrawing, formatting, printing or otherwise modifying the graphics output currently appearing on the screen. An example from the **Ordin** module is:



4.1.4 Graphics settings

If you wish to save graphics settings (colors, fonts, etc.) chosen in a particular run of a SYN-TAX module, then check the box **Save graphics settings** in the **Utilities/General options** dialog box. Its advantage is that you do not have to set these parameters when you start the program again. The settings are saved in the *prgname.INI* file, located in the same directory as the program itself. By deleting this INI file, the default graphic settings will become valid again upon the first call to the program.

4.2 Types of graphics

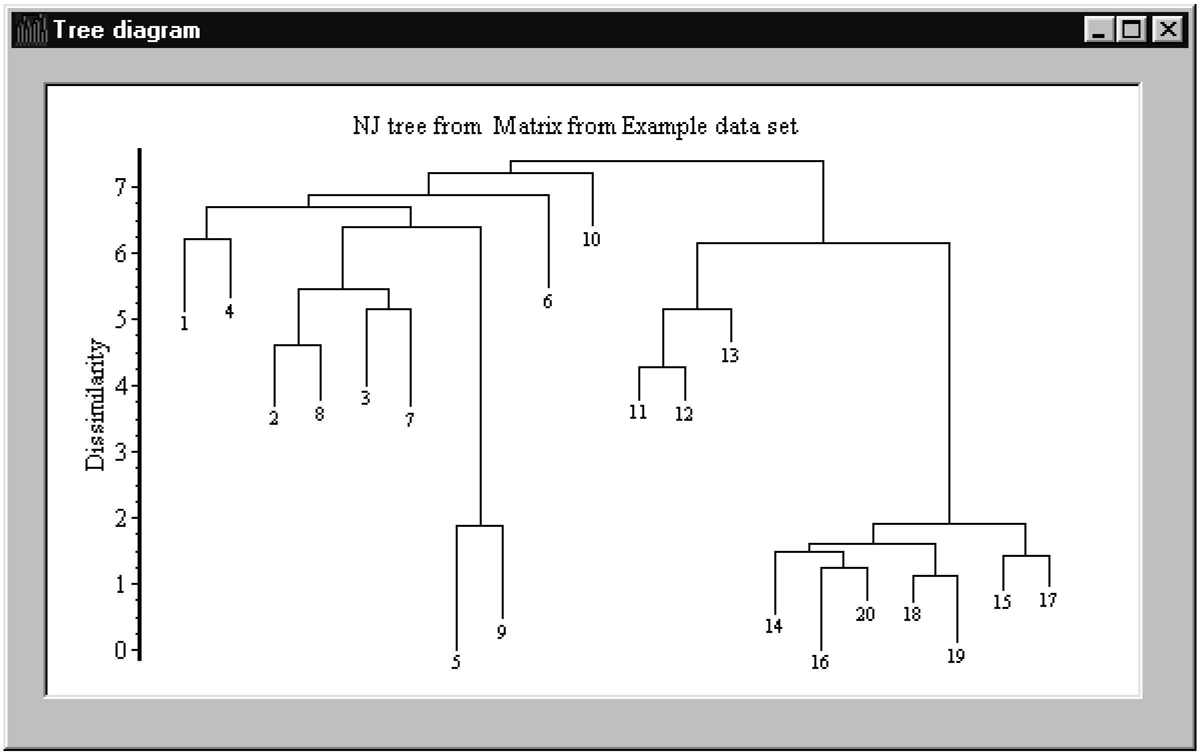
The main graphical objects displayed by SYN-TAX are discussed below in alphabetic order.

4.2.1 Additive trees

Additive trees are produced by neighbor joining analysis in SYN-TAX 2000. In these trees, the sum of branch lengths along the path between any two objects approximates the original distance between these two objects. There are two forms of additive trees:

- Rooted additive trees appear if the outgroup rooting or midpoint rooting option is selected for neighbor joining. The default extension for such tree files is .ADT;
- Unrooted additive trees appear otherwise. The default extension of these is .UTR.

The tree is displayed in a new graphics window when the **Draw tree** button is pressed after the analysis. For the example presented in the distribution disk, the rooted additive is displayed as



Additive trees can be reproduced from tree save files using the **Graphics/Draw rooted additive tree** or the **Graphics/Draw unrooted tree** commands, or by pressing, the following speed buttons

for a rooted additive tree, and

for an unrooted additive tree.

> **Note**: Unlike in minimum spanning trees, the interior noda of an additive tree DO NOT correspond with the study objects. The unrooted tree may be displayed showing its topology only, or showing branch lengths as well (see Section 4.3).

4.2.2 Biplots

Biplots are typically graphical superpositions of PCA object and variable scores. The variable scores are rescaled by an arbitrary scaling factor. Arrows drawn from the origin point to variable positions.

Biplots can be redrawn from exisiting files using the **Graphics/Draw biplot...** command or the speed button:

> **Note**: Arrows always appear in this diagram and rescaling is always in effect, even if the Object coordinates or the Variable coordinates were originally saved by some other ordination procedure. If

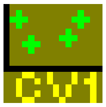
you do not want arrows and rescaling, use the **Joint plot** drawing facility (4.2.6).

4.2.3 Canonical correlation analysis ordinations

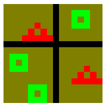
Canonical correlation analysis (COR) in SYN-TAX 2000 produces two save files for coordinates. They differ from other types of ordination score files in that only the first TWO dimensions are saved, and this is done automatically.

Object coordinates are saved such that scores for the canonical variates derived from the left set of original variables are arranged in the first two columns of the output file. Scores from the right set of variables are arranged in the third and fourth columns of the same file. That is, the four columns correspond to LCV1, LCV2, RCV1 and RCV2, respectively. When scattergrams are displayed after CANOCOR is performed, the title of the axes follows this convention. Therefore, if you want to see RCV1 and RCV2, then in the **New axes** pop-up dialog box, select axes 3 and 4! Furthermore, if you wish to see LCV1 and RCV1, then select axes 1 and 3!

CANOCOR ordinations saved previously in .ORD files can be reproduced using the **Graphics/Draw scatter for CANOCOR objects...** command, or the associated speed button:



Variable coordinates of CANOCOR are saved in the same way as variable coordinates for RDA, and CCoA results, that is, the two subsets of variables are distinguished by specifying the number of left set and right set variables in the output file. If you wish to reproduce a scattergram such that the two sets of variables are distinguished in the scattergram as well, use the **Graphics/Draw scatter for two sets of vars..**. command, or press the associated speed button:



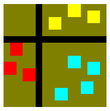
Hint: If you do not want to make such a distinction between the groups of variables, then display the coordinates from the same file using the standard **Graphics/Draw simple scattergram...** command.

4.2.4 Canonical variates analysis ordinations

Canonical variates analysis (discriminant analysis) in SYN-TAX 2000 produces two save files for coordinates, one for the objects and the other for the variables, with the extension .ORD.

In the file for object coordinates, the first line is a title, the second line contains the number of objects, the number of dimensions, the number of groups, and group size values. Then, in the subsequent lines, follow the coordinates. The file is completed by a line containing radii of confidence circles for the groups. The variable score file has a title in the first row, then the number of variables, the number of dimensions and the number of groups are specified. The coordinates are listed in the subsequent lines.

CANOVAR object ordinations saved previously can be reproduced using the **Graphics/Draw scatter for CANOVAR objects...** command, or the associated speed button:



In this way, the different groups are displayed in different colors in the ordination diagram. These colors are set automatically by the program so as to avoid the use of the background color for symbols as well. The default sequence of colors is as follows:

```
Red, Blue, Black, Yellow, Green, Gray, Fuchsia, Teal, Navy, Maroon, Lime, Olive,
Purple, Silver, Aqua, White
```

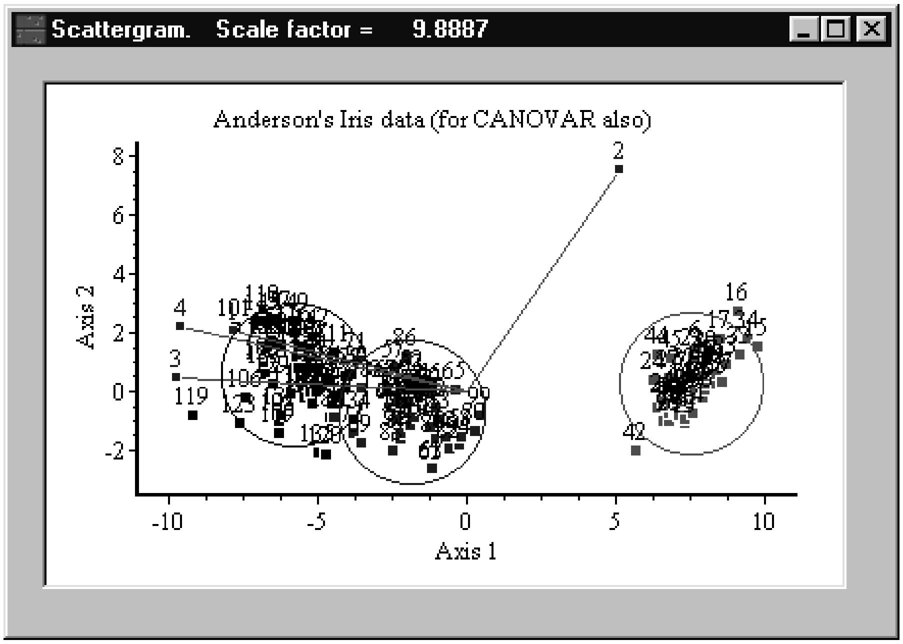
If there are more than 15 groups, then randomly generated colors are also used.

> **Note**: In CANOVAR scattergrams, the color for object symbols selected in the **Graphics/Symbol definition** dialog box is invalid.

If you do not want to see the points in different colors, reproduce the diagram from the same file using the **Graphics/Draw simple scattergram** command or the associated speed button.

CANOVAR biplots portray the object scatter and the variable positions in the same coordinate system, with variable coordinates arbitrarily rescaled. Arrows point towards the variable positions from the origin. The 5% confidence circles are drawn around the centroid of each group. To show the isodensity circles instead, which contain 95% of the individuals belonging to each group, use the respective pop-up menu command for the graphics. To reproduce the biplot from files saved previously, use the **Graphics/Draw CANOVAR biplot** command, or the associated speed button.

A graphics window example showing a CVA biplot for the *Iris* data set is


Scattergram. Scale factor = 9.8887
Anderson's Iris data (for CANOVAR also)
Axis 2
8
6
4
2
0
-2
-10
-5
0
5
10
Axis 1
2
4
3
119
16
42

4.2.5 Dendrograms

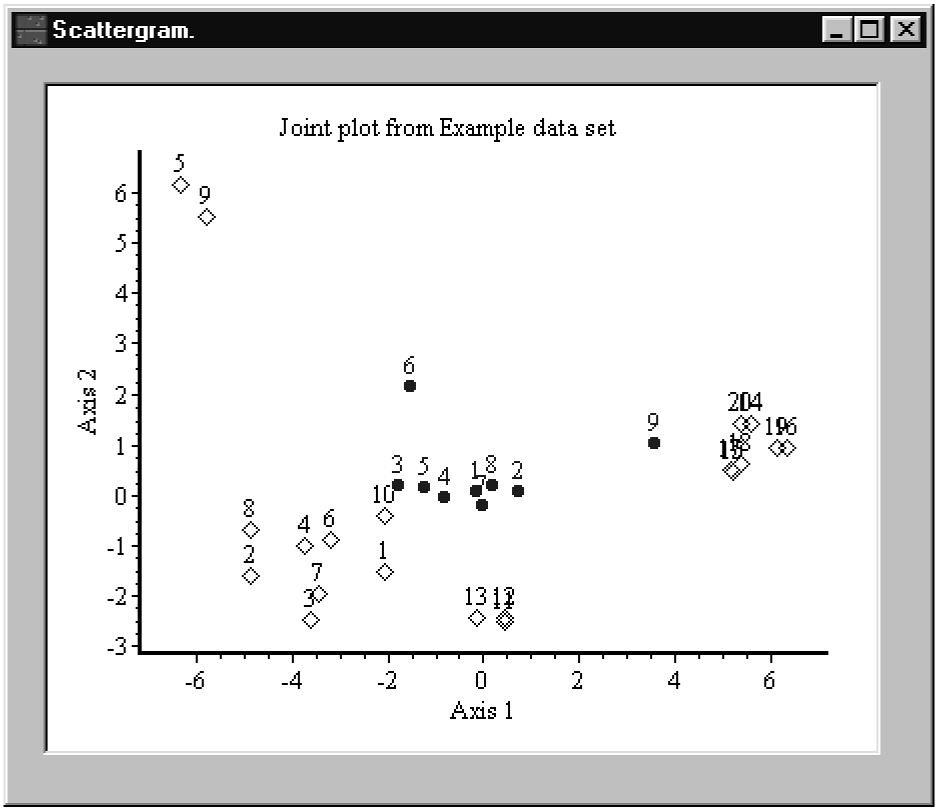
These are tree graphs typically displayed by hierarchical clustering methods in a new graphics window when the **Draw tree** button is pressed after the analysis. A dendrogram can be reproduced from a dendrogram save file by selecting the **Graphics/Draw dendrogram...** command from the Main Menu or by pressing the speed button:

The default extension of dendrogram save files is .DEN.

4.2.6 Joint plots

Joint plots are graphical outputs of correspodence analysis for the simultaneous display of object and variable ordinations. The relative positions of objects and variables are not arbitrary, i.e., there is no *a posteriori* rescaling of variable coordinates nor arrows as in PCA. The coordinates depend greatly on the weighting system, as selected in the submenu of the **Method/Correspondence analysis** menu. Joint plots can be reproduced from existing object and variable score files by the **Graphics/Draw joint plot...** command or the corresponding speed button:

A sample joint plot display is:



4.2 7 Minimum spanning trees

Minimum spanning trees are produced by SYN-TAX **HierClus** if the **Methods/Minimum spanning trees** option is chosen for an analysis. The tree is displayed in a new graphics window when the analysis is completed. A minimum spanning tree can be reproduced from a save file by the **Graphics/Draw min. spanning tree...** command or the corresponding speed button:

The default extension of minimum spanning tree save files is .MST.

Note: In minimum spanning trees, contrary to additive trees, all noda correspond to study objects. The sum of branch lengths in such a tree is the minimum. The tree may be displayed showing its topology only, or showing 'true' branch lengths as well, using the pop up or **Graphics** menu command **Unrooted tree shape**.

4.2.8 Rotating plots: 3D scattergrams

On the plane only two dimensions can be portrayed effectively. There is, however, a procedure which gives you the illusion of three dimensions, the rotating plot technique first programmed by Fisherkeller, Friedman and Tukey in 1972. The basic idea is that a three dimensional coordinate system together with the points are rotated continuously and their two dimensional projection is shown on the screen. In this way we have a continuously changing view on the shape of the point cloud, and we can find optimal and most interesting orientations. The rotation axis is a horizontal line in the middle of the screen; the current angles of the x, y and z axes to this axis are displayed on the right side.

In order to use this facility, which is the only DOS routine remaining in SYN-TAX 2000, the ordination score or the raw data file must have at least 3 columns (dimensions). The data may be simple, or grouped by objects. In the latter case, groups will appear in different color on the screen.

The program reads a data matrix with dimensions as columns (no upper limit) and points (objects) as rows (max. 500). The user defines the axes for the starting display in a DOS window, and then the **D** key may be used to select other dimensions (or even the same original axis for x, y and z). Further hot keys are:

A (or a): to hide or show the axes;

N (or n): to add or remove id. numbers;

x, X, z and Z: to change angle of X, Y and Z axes to the rotation axis (case of letter influences direction of changing the angle). With some experience you will be able to find any orientation of the point cloud examined.

T (or t): to change the direction of rotation;

S (or s): to stop rotation and to allow for stepwise motion;

C (or c): to restart continuous rotation;

R (or r): to record the coordinates of the actual projection in file PROJ*, where * is 1,2,3,...,99. The saved values are not affected by the zoom. These files must be renamed if needed for future use.

Esc: terminates the program, control returns to WINDOWS;

F1 and F2: to change speed of rotation (shown on top left);

F8: Save PCX file of the image (the default name of the file saved appears on the screen);

F9: Save TIF file of the image (the default name of the file saved appears on the screen);

PgUp and PgDn: change increment of rotation (in radians, shown on right). If this value reaches zero, the rotation is arrested.

I (or i): zoom in = enlargement in size;

O (or o): zoom out = reduction in size;

Arrow keys (←, ↑, →, ↓): shift the whole diagram.

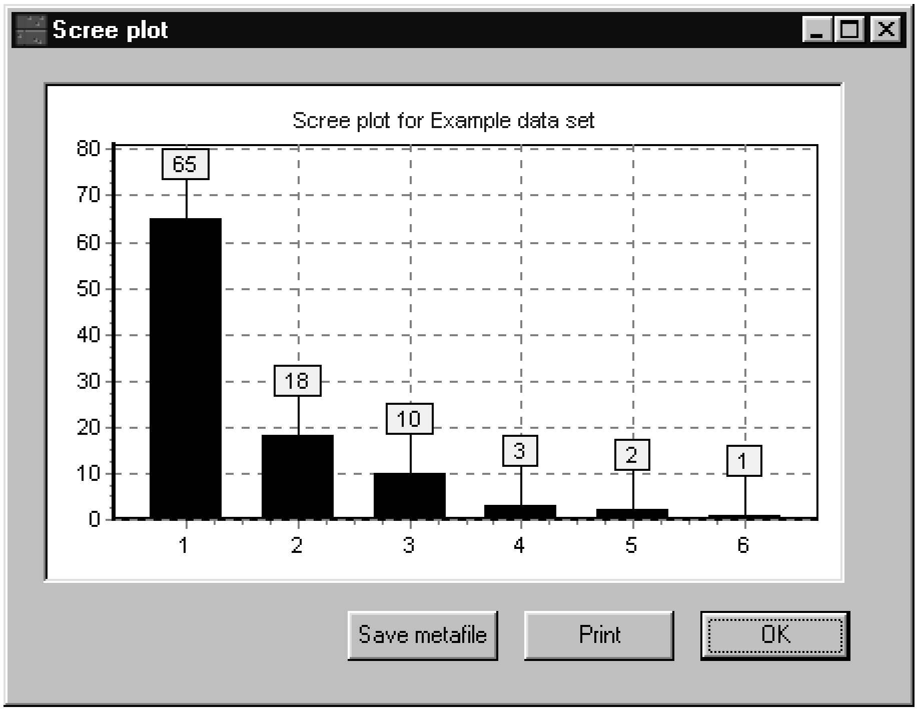
The faster your processor and higher the resolution of your monitor, the better the illusion of three dimensions. Careful combination of **PgUp, PgDn** and the **F1-F2** keys may be used to adjust the virtual speed of point movements on the screen, thus finding an optimal speed on your system.

Save screen images: TIF or PCX with default names. The files can be renamed after execution of the rotating plot routine.

The program will use different colors if the data points are grouped (up to 15 groups, in this case rows of the data matrix must be grouped accordingly, as described in Section 3.1.3). This program can be called from the **Utilities/Graphics** menu or by pressing the speed button:

4.2.9 Scree plots

When a metric ordination is completed, a diagram showing the percentage importance of ordination axes (the associated eigenvalues) is displayed if the **Scree plot** button is pressed. The maximum of eigenvalues shown is 20. The percentages are rounded to the nearest integer, for more precise values, see the **Summary** of results. A scree plot has the following appearance in SYN-TAX:



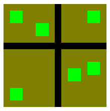
Scree plots cannot be reproduced from files.

4.2.10 Shepard diagrams

After nonmetric multidimensional scaling (NMDS), Shepard diagrams may be displayed to show the relationship between ordination and original distances. The regression 'line' showing the efficiency of monotonic regression may be added to or removed from the diagram using the **Remove/Add regr. line** button. Note that Shepard diagrams cannot be reproduced from files.

4.2.11 Simple scattergram

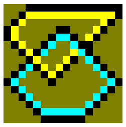
Use this facility if you wish to display a two-dimensional ordination of a single set of objects without grouping. The speed button is:



4.2.12 Superimposed convex polygons

Ordinations and partitions may be compared graphically by superimposing the clusters as convex polygons on an ordination scattergram. On the ordination plane each cluster is represented by the smallest convex polygon that can be drawn around the members of that cluster. These polygons are shown in different color.

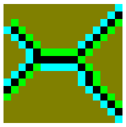
Use the **Graphics/Superimpose convex polygons** command in **Ordin** or **NonHier**, or click the speed button:



Then, specify an input filename for the partition (as a group membership vector) and a filename for the ordination file. You may wish to use labels for the objects as well, otherwise cancel the Open labelfile dialog box. Then, the diagram appears in a graphics window. You may want to try the **Remove/Add polygons** command in the pop up menu. This is useful if you want to show a partition using only the colors.

4.2.13 Superimposed minimum spanning trees

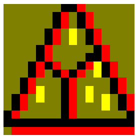
A minimum spanning tree superimposed over an ordination has many advantages in clarifying ordination scatters. Use the **Graphics/Superimpose minimum spanning tree** command or the corresponding speed button in **Ordin**:



Then, specify a filename for the tree and another file for the ordination scores in the Open dialog box. You may wish to use labels for the objects as well, otherwise cancel the labelfile dialog box.

4.2.14 Ternary plots

When the number of clusters is three in fuzzy clustering, ternary plots can be drawn. Such a plot is a so-called simplex, illustrating the grouping tendency of each object. The tips of the triangle represent the three clusters, and the closer a point to a given tip, the higher its association to the group this tip represents. If an object is positioned in the centroid of the triangle, then it has cluster membership values of 0.3333 for all the three clusters. Ternary plots can be reproduced from existing fuzzy classification score files if the number of clusters is 3, by the **Graphics/Draw ternary plot..**. command or the associated speed button:

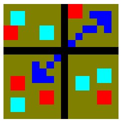


> **Note**: if a fuzzy classification has two, or more than three clusters, then these diagrams cannot be displayed. In these cases, an ordination-like scattergram may provide an illustration.

To change the color of the triangle, use the **Triangle color** command.

4.2.15 Triplots

Triplots are graphical displays typical of RDA and CCoA results. The left and right set of variables are distinguished from each other, and arrows point to the positions of right set (constraining) variables. Triplots can be reproduced from existing ordination and variable score files by the **Graphics/Draw triplot...** command or the associated speed button:



> **Hint**: If you wish to display only the ordination of variables from a file, then use **Graphics/Draw scatter for two sets of vars..**. command, or press the corresponding speed button.

4.3 Graphics commands

Graphics commands are available directly from the **Graphics** menu of the **Main Window**, from the pop up menu when a graphics window is open, or from both. These are useful to modify the appearance of the diagram, or to print or copy the graphics, and so on. The commands are discussed in alphabetic order.

4.3.1 Axis caption font

This font determines the style of the caption to the axes and the legend to the units.

> **Important:** If the caption on the vertical axis is horizontal, then select a TrueType font in the **Graphics/Axis caption font** menu. Fonts other than this type cannot be rotated.

4.3.2 Axis color

This command can be used to modify the color of the horizontal axis in dendrograms or both axes in ordination displays. Use the **Axis caption font** command if you want to modify the color of the caption as well.

4.3.3 Background color/Panel color

The **Background color** command, which calls the standard Color dialog box, is used for specifying color of the entire graphics screen for tree diagrams. In ordinations, the background color is used only for the ordination plane, while the area outside the coordinate system may be controlled by the **Panel color** command.

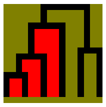
4.3.4 Copy

The **Copy** command in the pop-up menus can be used to copy the image into the Clipboard for use by word processors or graphics programs. The format is .BMP.

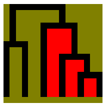
4.3.5 Dendrogram shape

The program offers three alternative views of the same dendrogram:

1) At each node the larger group is placed on the left;

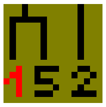


2) At each node the larger group is placed on the right;



and

3) At each node the group with the smaller id. number is placed on the left (default)



The user may switch between these modes using the pop-up menu command **Dendrogram shape**, or using the same command available from the **Graphics** Menu. After a new option is chosen, the actually displayed tree changes its shape immediately.

4.3.6 Dragging diagrams

The entire ordination scattergram or tree may be repositioned by holding down the right mouse button and dragging the configuration into the desired direction. It is especially useful if one wishes to reposition the point scatter for optimal fit to the graphics area. Observe that the axes and their captions also change during this operation.

4.3.7 Labels/Labeling objects/Labeling variables

The points prepresenting objects, or sometimes variables, can be labeled in SYN-TAX graphics displays in several ways. If no labelfile is open, then the points are labeled by numbers. If a labelfile was opened before the graphics is displayed, then **Text** labels may be used to identify the points. Use the commands in the **Graphics** menu or the pop up menu to modify labeling.

You can remove both numbers and text labels. In this case, dendrograms will be completely unlabeled, whereas in unrooted trees and ordinations symbols will identify the points. To modify the appearance of symbols, use the **Symbol definition** command.

In ordination scattergrams, labels can exceed 8 characters in length. For trees, there is a limit of 8 characters, so that the first 8 characters are retained if a label is longer. Trailing and leading spaces are truncated.

For dendrograms, the **Labels** command may be used to switch between horizontal and vertical orientation. This is especially useful if long horizontal labels overlap in large dendrograms. For more on labels, see Section 3.5.

4.3.8 Label font/Object label font/Variable label font

The style, size and color of labels, either text or number, can be modified using the standard font dialog.

4.3.9 Line width/Printer scaling

The width of lines (e.g., tree branches, biplot arrows) can be modified by the user. The value is given in pixels.
This command, available from the **Graphics** menu or from the graphics pop-up menu, can be used to control the printed graphics output. When scaling is 100%, the diagram will take a full page, its actual size depending on whether Landscape or Portrait orientation is selected in the **Printer Setup** dialog box. Lower percentages will produce a reduced image such that font sizes and line width are NOT affected. If you wish to print a reduced image such that fonts are also small, change font size before printing or use the zoom procedure. The value of scaling does not influence the display on the monitor.

4.3.10 New axes

When displaying ordination diagrams, the **New axes** pop up menu command can be used to select a new combination of horizontal and vertical axes. In the header of the dialog box, the user is informed about the number of axes actually available in the input file. Entering a larger number is not allowed, an error message appears if nevertheless attempted.

If you want to display an axis that does not exist in the ordination score file, then the analysis has to be repeated with a larger number of axes. To specify the number of axes, use the **No. of axes** box in the **Main Window** of **Ordin**.

4.3.11 Print

This command is available from the pop-up menu when a graphics window is active. By clicking this menu item the diagram is sent to the currently selected printer. The scaling of the diagram and line width in pixels can be modified previously using the **Line width/Printer scaling** command from the same pop-up menu or from the **Graphics** menu. Note that the printer can only be selected through the **Printer setup** command in the **File** menu.

4.3.12 Resizing graphics

There are two ways to change the size of the graphics window. **Zoom** or **Stretch** provides you an automatic procedure. However, if you wish to do resizing manually, use the right mouse button and keep it pressed on the lower right corner of the graphics window until the desired window size is reached. Finally, choose the **Resize to fit** command from the pop-up menu of the graphics and the diagram will attain the desired dimensions.

> **Hint:** If you want to see how the diagram modifies when resizing the window, check the **Automatic resize** box in the **Utilities/General options** dialog box beforehand. Given this box checked, the diagrams are redrawn upon the slightest modification of the size of the Graphics Window.

4.3.13 Save

This command is available from the pop-up menu when a graphics window is active. By clicking this menu item, the drawing is saved in BMP format in a file specified by the user in the **Save Files** dialog box.

4.3.14 Scaling axes

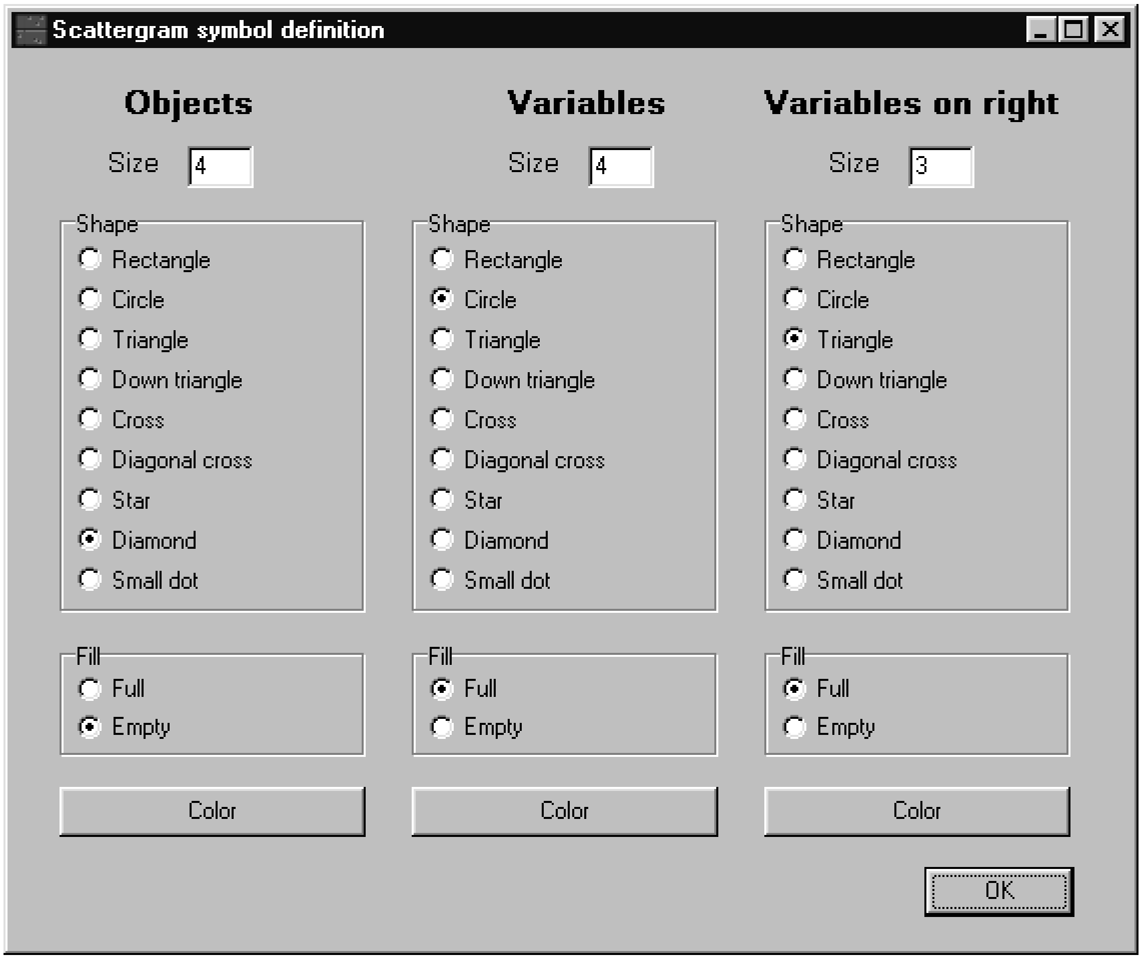
In ordinations, the unitsare normally of the same physical size on both axes, called **proportional** scaling. It may happen that in proportional scaling one axis becomes too short, however. In order to display an ordination utilizing the entire graphics window, use **Not proportional** scaling of axes.

4.3.15 Symbol definitions

In unrooted trees and ordinations, the objects may be represented by symbols. The shape, size and color of these symbols may be changed using the **Symbol definitions** command available from the **Graphics** menu or from the pop up menu. In scatter diagrams, labels or id. numbers - if used - appear on the top of the symbol.

> **Note**: For ordinations with grouped objects, the empty symbol option does not work.
> **Hint:** If you do not want to see symbols in an ordination scattergram, only labels or id. numbers of objects, then select 1 as symbol size, and set symbol color to be the same as the background color. Note that 0 size cannot be set, it causes a graphics error!

The following dialog box appears when there is a triplot, i.e., objects and two types of variables, in the graphics window of **Ordin**:



Scattergram symbol definition

Objects	Variables	Variables on right
Size 4	Size 4	Size 3
Shape	Shape	Shape
○ Rectangle	○ Rectangle	○ Rectangle
○ Circle	● Circle	○ Circle
○ Triangle	○ Triangle	● Triangle
○ Down triangle	○ Down triangle	○ Down triangle
○ Cross	○ Cross	○ Cross
○ Diagonal cross	○ Diagonal cross	○ Diagonal cross
○ Star	○ Star	○ Star
● Diamond	○ Diamond	○ Diamond
○ Small dot	○ Small dot	○ Small dot
Fill	Fill	Fill
○ Full	● Full	● Full
● Empty	○ Empty	○ Empty

In other cases, the "Variables on right" section of this dialog box is inactive.

4.3.16 Title font

The title is a string of characters appearing in graphics windows. Its font may be changed by the standard font dialog. If you wish to change the text as well, type the new text into the **Title** box of the **Main Window**, and use the **Resize to fit** command. The diagram will be redisplayed with the new title on the top.

4.3.17 Tree color

The color of tree branches may be changed by this command, which uses the standard Color dialog.

4.3.18 Triangle color

In ternary plots, appearing when a fuzzy classification contains three clusters, the color of the triangle can be modified by the standard Color dialog.

4.3.19 Unrooted tree shape

In graphics displays with a minimum spanning tree or an unrooted additive tree on screen, there are two possibilities:

1) True tree lengths are not depicted, only tree **topology** is illustrated by branches of arbitrary and fairly uniform length;

2) The length of branches is proportional to the original length values, so that the user has more ideas about within-tree distances than in the previous case, but the labels may overlap considerably in many situations. This is a new and unique feature of SYN-TAX.

4.3.20 Zoom

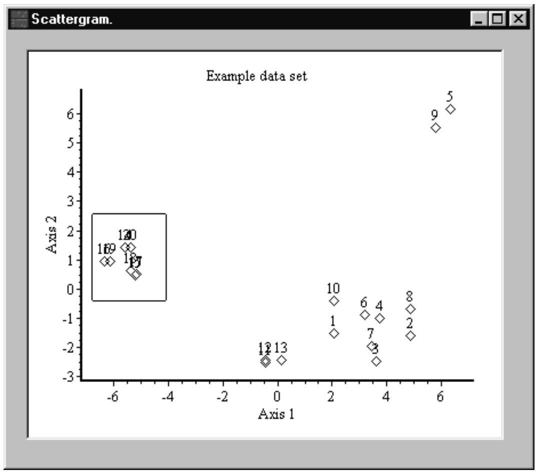
There are three kinds of zooming graphics in SYN-TAX 2000.

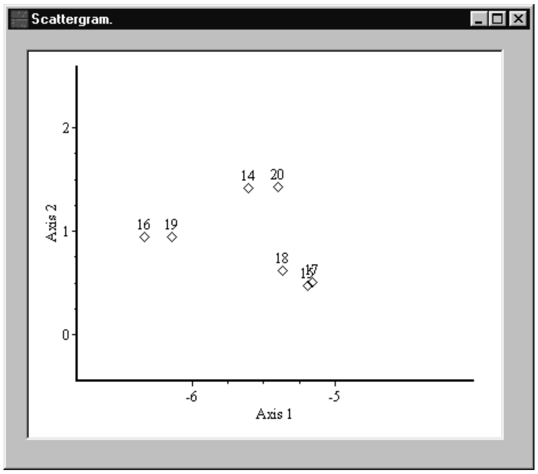
1) Some diagrams can be zoomed using the pop-up menu command **Zoom+ (in)** or **Zoom- (out)**. The first produces an enlarged full diagram, whereas the second command reduces diagram size step by step. Scroll bars appear or disappear when necessary. **Zoom+** is most useful before printing large diagrams: font size is scaled down when the diagram is sent to the printer port. In other words, the entire zoomed diagram is reduced in size when printed.

2) Zooming in horizontal direction only (**Stretch**) is useful to display and print very large dendrograms such that the labels do not overlap.

3) Pressing the *left* mouse button and dragging a rectangle from top left to bottom right over any portion of the diagram will provide an enlargement of the selection. Then, dragging any rectangle in the opposite direction will reproduce the full diagram. This option is very useful to examine small details of the results, especially when labels overlap in ordination scatterplots of many objects.

This second zooming option is exemplified below. The window on left shows the full ordination diagram, with the zooming rectangle. Note the overlap of labels and symbols. After releasing the left mouse button, the selected rectangular area is magnified in the same window (right), allowing the inspection of labels. Within the selection, you can do further zooming if some labels still overlap to some extent.


Scattergram.
Example data set
Axis 2
Axis 1
6
5
4
3
2
1
0
-1
-2
-3
-6
-4
-2
0
2
4
6
5
9
14
20
16
19
18
15
17
10
6
4
8
1
2
7
3
11
12
13


Scattergram.
Axis 2
Axis 1
2
1
0
-6
-5
14
20
16
19
18
15
17

Chapter 5

Theory in brief

5.1 Data transformation and standardization

There are two ways of modifying the original data in SYN-TAX:

- run-time standardization/ transformation using the **Standardization** menu item in the Main Window, and
- permanent standardization/transformation using the **Utilities/Data standardization** command before starting any data analysis.

In the first case the transformed data are not saved, the original data remain intact, and there are only six options to modify the data (numbers below in the table). In the second case, a new file with the transformed data is saved, the number of options is about 20.

	Name	Formula (x_{ij} is the original value, x'_{ij} is the transformed value)
1	St. by standard deviation of rows	$x'_{ij} = (x_{ij} - \overline{x}_i) / s_i$
	St. by standard deviation of columns	$x'_{ij} = (x_{ij} - \overline{x}_j) / s_j$
2	St. by range of rows	$x'_{ij} = (x_{ij} - \min_i \{x_{ij}\}) / (\max_i \{x_{ij}\} - \min_i \{x_{ij}\})$
	St. by range of columns	$x'_{ij} = (x_{ij} - \min_j \{x_{ij}\}) / (\max_j \{x_{ij}\} - \min_j \{x_{ij}\})$
3	Logarithmic transformation ($x \geq 0$, $c > 1$)	$x'_{ij} = \log_c (x_{ij}+1)$
4	Clymo's transformation of percentages ($100 \geq x \geq 100$; $c \neq 0$)	$x'_{ij} = (1 - e^{(-cx_{ij})}) / (1-e^{-c})$
5	Power (exponential) transformation ($c>0$)	$x'_{ij} = x_{ij}^{(1/c)}$
6	Arc sin transformation ($-1 \leq x \leq 1$)	$x'_{ij} = \text{arc sin sqrt} (x_{ij})$
	Division of values by row total	$x'_{ij} = x_{ij} / \sum_i x_{ij}$
	As above multipled by 100	$x'_{ij} = 100\, x_{ij} / \sum_i x_{ij}$
	Division of values by column total	$x'_{ij} = x_{ij} / \sum_j x_{ij}$
	As above multiplied by 100	$x'_{ij} = 100\, x_{ij} / \sum_j x_{ij}$
	Division by row maximum	$x'_{ij} = x_{ij} / \max_i \{x_{ij}\}$

Division by column maximum	$x'_{ij} = x_{ij} / \max_j \{x_{ij}\}$
Centring about row means	$x'_{ij} = (x_{ij} - \bar{x}_i)$
Centring about column means	$x'_{ij} = (x_{ij} - \bar{x}_j)$
Double centring	$x'_{ij} = x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}$
Binarization	$x'_{ij} = 0$ if $x_{ij} = 0$, $x'_{ij} = 1$ if $x_{ij} > 0$
Adjustment to unit row length	$x'_{ij} = x_{ij} / \sqrt{\sum_i x_{ij}^2}$
Adjustment to unit column length	$x'_{ij} = x_{ij} / \sqrt{\sum_j x_{ij}^2}$

Implicit binarization: if a presence/absence coefficient is selected for clustering or ordination, and your data file contains counts, percentage cover etc., then the program will automatically convert the data into binary form. All zeros remain zeros, while all positive values become 1.

> **Hint**: You can use a combination of two or more methods by repeated transformations, for eample, a log transform first, and then some other. The effects of combined transformations may be more difficult to understand, however.

5.2 Distance and dissimilarity

Most routines of SYN-TAX 2000 use distances and dissimilarities to measure resemblance between objects. When the selected Method requires, the **Coefficient** menu in the Main Window becomes activated. Within this menu, coefficients are grouped according to the scale type they use: binary data (presence/absence), ordinal data, ratio scale data and mixed data. These submenus are activated in a context-dependent manner.

Similarity and correlation measures are transformed into dissimilarity form, according to the following:

$d_{jk} = 1 - s_{jk}$ (similarity into dissimilarity)

There are some exceptions, because principal components analysis and canonical correlation analysis use untransformed correlations (or covariances, cross products). In these cases these measures are applied to variables, however. Principal coordinates analysis transforms the measures into distances, when necessary.

Abbreviations:
x_{ij} is the score for variable i and object j
a+b+c+d=n, in the 2x2 contingency table
n is the number of variables
m is the number of objects

5.2.1 Coefficients for binary (presence/absence) data

16 coefficients are available.

PHI	$1-[(ad)-(bc)] / \sqrt{[(a+b)(a+c)(b+d)(c+d)]}$

Yule	$1-[(ad)-(bc)] / [(ad)+(bc)]$
Jaccard	$1-a / (a+b+c)$
Simple matching	$1-(a+d) / n$
Russell & Rao	$1-a/n$
Sorensen	$1-2a / (2a+b+c)$
Rogers & Tanimoto	$1-(a+d) / (a+2b+2c+d)$
Baroni-Urbani & Buser	$1- [\sqrt{(ad)}+a] / [\sqrt{(ad)} +a+b+c]$
Ochiai	$1-a / \sqrt{[(a+b)(a+c)]}$
Anderberg #1	$1- \sqrt{[(a/(a+b)) (a/(a+c)) (d/(b+d)) (d/(c+d))]}$
Anderberg #2	$1-[(a/(a+b))+(a/(a+c))+(d/(b+d))+(d/(c+d))] / 4$
Kulczynski (symm.)	$1-[(a/(a+b))+(a/(a+c))] / 2$
Sokal & Sneath	$1-a / (a+2b+2c)$
Chord d. (binary case)	$\sqrt{2[1-a / \sqrt{((a+b)(a+c))}]}$
Faith	$1 - (a + d/2) / n$
Euclidean d. (binary case)	$\sqrt{(b+c)}$

For a fuller discussion of these coefficients, see Podani (2000).

> **Note:** If a presence/absence coefficient is selected for clustering or ordination, and your data file contains counts, percentage cover etc., then the program will automatically convert the data into binary form. All zeros remain zeros, while all positive values become 1.

5.2.2 Coefficients for ordinal data

Resemblance coefficients offered by the previous SYN-TAX versions accepted 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 tau for objects *j* and *k* is given by

$$\tau_{jk} = 2\,(a - b) \;/\; \sqrt{[n(n-1) - 2T_j]\;[n(n-1) - 2T_k]}$$

where n is the number of variables, a is the number of pairs of variables ordered for objects j and k identically, b is the number of pairs of variables that are reversely ordered in j and k. T_j and T_k are the numbers of tied variable pairs in objects j and k, respectively.

Goodman and Kruskal's gamma is simpler,

$$\gamma_{jk} = (a - b) / (a + b)$$

A new *measure of discordance* (Podani 1997b) may be recommended when presence/absence is also meaningful (e.g. phytosociological data with Braun-Blanquet scores):

$$DC_{jk} = 1 - 2(a - b + c - d) / [n(n-1)]$$

where n, a and b are defined as above. c is the number of pairs of variables tied in both j and k, corresponding to joint presence or joint absence, as in the examples given below

```
1 1  or  1 2  or   0 0
1 1      1 2       0 0
```

That is, such pairs of variables (rows) increase the similarity (decrease the dissimilarity) of the objects. d is the number of all pairs of variables that are tied at least for one of the objects being compared such that either one, two or three scores are zero. The following examples will clarify this:

```
1  0   or   1   1     or    1   0   or   0  1
0  0        1   0           1   0        0  3
```

These pairs of variables indicate contradiction of the objects at least in presence/absence relations and will contribute to increased dissimilarity.

The τ and γ functions are provided as *complements*, i.e. in dissimilarity form. These coefficients are available through ordinal non-hierarchical or ordinal hierarchical clustering and non-metric multidimensional scaling. If you wish to use them 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: the *ranks of the dissimilarity values* may be saved as a separate matrix. You may want to do that if you wish to perform non-hierarchical ordinal clustering as well, since the time-consuming ranking process can be skipped in the second analysis.

5.2.3 Coefficients for ratio scale data

The available coefficients are:

Correlation	$1 - \sum_i (x_{ij} - \overline{x}_i)(x_{ik} - \overline{x}_k) / \; / \sqrt{[\sum_i (x_{ij} - \overline{x}_i)^2 \sum_i (x_{ik} - \overline{x}_k)^2]}$
Bray - Curtis	$1 - 2 \sum_i \min\{x_{ij}, x_{ik}\} / \sum_i \{x_{ij} + x_{ik}\}$
Ruzicka	$1 - \sum_i \min\{x_{ij}, x_{ik}\} / \sum_i \max\{x_{ij}, x_{ik}\}$
Similarity ratio	$1 - \sum_i x_{ij} x_{ik} / (\sum_i x_{ij}^2 + \sum_i x_{ik}^2 - \sum_i x_{ij} x_{ik})$
Horn	$1- [\sum (x_{ij}+x_{ik})\log(x_{ij}+x_{ik}) - \sum x_{ij}\log x_{ij} - \sum x_{ik}\log x_{ik}] /$

$[\ (x_{.j}+x_{.k})\log(x_{.j}+x_{.k}) - x_{.j}\log x_{.j} - x_{.k}\log x_{.k}\ ]$

City block metric $\sum_i |x_{ij}-x_{ik}|$

Mean character difference $\sum_i |x_{ij}-x_{ik}| / n$

Canberra metric $\sum_i |x_{ij}-x_{ik}| \ / \ (|x_{ij}| + |x_{ik}|)$

Normalized Canberra metric $\{ \sum_i |x_{ij}-x_{ik}| \ / \ (|x_{ij}| + |x_{ik}|) \} / n$

Euclidean distance $\sqrt{\sum_i (x_{ij}-x_{ik})^2}$

Chord distance $\sqrt{2\ [\ 1-\Sigma_i\ x_{ij}x_{ik} \ / \ \sqrt{(\sum_i x_{ij}^2 + \sum_i x_{ik}^2}\ )\ ]}$

Angular separation $1 - [\sum_i x_{ij}x_{ik} \ / \ \sqrt{(\sum_i x_{ij}^2 + \sum_i x_{ik}^2}\ )\ ]$

Penrose size (HierClus only) $\sqrt{1/n^2\ [\ \sum_i (x_{ij}-x_{ik})\ ]^2}$

Penrose shape (HierClus only) $\sqrt{1/(n-1) \sum_i (x_{ij}-x_{ik})^2 - 1/n(n-1)\ [\ \sum_i (x_{ij}-x_{ik})\ ]^2}$

Balakrishnan - Shangvi $\sqrt{\sum_i [\ (x_{ij}-x_{ik})^2 / (x_{ij}+x_{ik})\ ]}$

Generalized distance* $D^2 = (\mathbf{x}_j - \mathbf{x}_k)\ \mathbf{W}^{-1}\ (\mathbf{x}_j - \mathbf{x}_k)$

where $\mathbf{x}_j$ and $\mathbf{x}_k$ are mean vectors and **W** is the pooled within-group covariance matrix.

Weighted dissimilarity $\sum_i p_i |x_{ij}-x_{ik}| \ / \ \sum_i p_i$; $p_i = \sum_j x_{ij} / m$

5.2.4 Coefficients for mixed data

Gower $1 - \sum_i w_{ijk}s_{ijk} / \sum_i w_{ijk}$

where $w_{ijk} = 0$ if comparison of j and k is
not possible for variable i (missing data);

and

a) for binary variables:

$w_{ijk} = 1$ and $s_{ijk} = 0$ if $x_{ij} \neq x_{ik}$

$w_{ijk} = s_{ijk} = 1$ if $x_{ij} = x_{ik} = 1$ or if

$x_{ij} = x_{ik} = 0$ and double zeros are included;

$w_{ijk} = s_{ijk} = 0$ if $x_{ij} = x_{ik} = 0$ and double zeros are excluded;

b) for nominal variables:

$w_{ijk} = 1$

$s_{ijk} = 0 \quad \text{if } x_{ij} \neq x_{ik}$

$s_{ijk} = 1 \quad \text{if } x_{ij} = x_{ik}$

c) for variables measured on the interval scale:

$w_{ijk} = 1$ and

$s_{ijk} = 1 - | x_{ij}-x_{ik} | \ / \text{(range of variable i)}$

Distance for mixed data $\sqrt{\ } \{ \sum_i w_{ijk} [(x_{ij} - x_{ik}) / q_{ijk}]^2 \}$

where $w_{ijk} = 0$ if comparison of j and k is not allowed for variable i (missing data)

a) for binary variables:

$q_{ijk} = 1$

b) for nominal variables:

$q_{ijk} = x_{ij} - x_{ik} \quad \text{if } x_{ij} \neq x_{ik}$

$q_{ijk} = 1 \quad \text{if } x_{ij} = x_{ik}$

c) for variables measured on the interval scale:

$q_{ijk} = \max (x_{ij}) - \min (x_{ij}); \ j=1,...m$

For both coefficients and,

d) for **ordinal** variables for all scores are replaced automatically by their ranks (r_{ij}), $w_{ijk} = 1$ and then we define

$$s_{ijk} = 1 - \frac{| r_{ij} - r_{ik} |}{\max \{r_i\} - \min \{r_i\}}$$

where the numerator is the minimum number of interchanges in the rank order needed to put an object with the same value as x_{ij} into an object with the same value as x_{ik}. The denominator is the possible maximum for variable i.
If ties do not occur, the formula implies relative rank differences. Ties can also be considered in the interchange metric.

In data files with mixed data the variables must be arranged in the following sequence:

- binary variables (coded with 0-s and 1-s);
- multistate (nominal) variables (coded with 0, 1, 2, ...);
- ordinal variables, arbitrary ordinal coding, treated by the interchange measure (see Podani 1989 in Taxon, for more);
- ordinal variables, arbitrary ordinal coding, treated by the relative rank difference (see Podani 1989 in Taxon, for more);
- variables measured on the interval or ratio scale ("quantitative variables").

An example with two binary variables, one multistate variable (with 4 states), one ordinal variable and two quantitative variables for 10 objects is this:

```
Sample file for mixed data
6  10
0 1 1 0 0 1 1 1 1 1
0 1 1 1 1 1 0 0 0 0
1 2 2 2 1 3 0 0 0 2
1 2 3 3 3 2 1 1 1 1
12.3 4.5 5.6 5.7 10.0 2.3 4.5 6.5 4.1 6.9
```

```
12.0 10.1 10.3 10.5 10.6 12.3 16.5 12.2 10.0 9.8
```

The user specifies runtime in a dialog box, appearing when a given mixed data coefficient is chosen in the **Coefficient** menu, the number of variables belonging to each scale type. In this dialog box it is to be also specified whether the double zeros are considered or disregarded.

Missing data are also allowed in these cases, each unknown data item must be coded by a negative dummy value (e.g., -1.0). It means that the original data can have only non-negative values, a condition easily satisfied.

5.2.5 A coefficient for nominal data

There is no explicit option for a coefficient for nominal data in SYN-TAX 2000. Nevertheless, if the data contain only nominal variables, you may choose the mixed data option and then specify all variables as nominal in the dialog box. Consider double zeros (important!), so that you have the simple matching coefficient generalized to nominal data:

d = 1 - (no. of variables in which the two objects agree) / n

5.3 Non-hierarchical clustering

In the classical sense, non-hierarchical clustering generates a (hard or crisp) partition of objects into p mutually exclusive groups (or clusters). Use the **NonHier** module to achieve this. Some methods require *a priori* specification of the number of clusters and start from an initial partition which is refined by relocations until a final optimum is reached. The number of clusters may be set in the **Number of clusters** box near the **Summary** button.

5.3.1 *K*-means clustering

It minimizes the sum of squares within clusters. The use of this strategy is limited to situations when quantitative raw data are available and the sum of squares criterion is meaningful (it cannot be used, for example, for nominal and ordinal data).

If this strategy is selected from the **Method** menu, then the small box labeled **Number of clusters** is activated near the **Analyze** button, so that the user can enter the desired value.The starting classification may be:

- Randomly generated partition.
- User specified partition (e.g., derived from a previous hierarchical classification using the **Partition from a dendrogram** command in the **Utilities** menu).
- The user provides seed objects, and then each object is linked with the closest seed object to form an initial partition.
- The program selects random seeds.
- The p objects that fall farthest apart in the n-dimensional space are selected as seed objects for the starting partition for p clusters.

Since the final partitions are usually not unique from random starts, you are advised to perform many analyses when the *Number of searches* dialog box appears, and then to accept the most optimal result.

5.3.2 Multiple partitioning

It requires specification of the maximum number of clusters, say p, into which the objects should be divided. The analysis starts with one single cluster which is subdivided into two groups by finding the object most distant from the initial centroid. In an iterative procedure the objects are assigned to the closest group, thus refining the clustering. When the two-cluster classification is stable, the analysis turns to the three-cluster classification by finding the object that falls farthest apart from its own centroid, and iterative relocations restart. The analysis stops when the *p*-cluster partition is stable. The method assumes that Euclidean distance is meaningful, other distance measures are not compatible with the clustering model. In some sense, the result is a hierarchical, although not necessarily dichotomous classification, based on k-means clustering at each level.

5.3.3 Quick clustering

This procedure was formerly recommended for very large problem sizes that exceeded the limits imposed by more sophisticated clustering methods. Although these limits are hard to reach in current hardware, the method is retained in SYN-TAX 2000 because it does not require *a priori* specification of the number of clusters and is therefore still useful for data exploration. What is to be defined, however, is an arbitrary distance or dissimilarity radius. In the first step, the program selects a random object and finds all others that are closer to this seed than the specified radius. They will form the first "quick" cluster. Then the next random object is selected which is not yet clustered, representing the seed of the second cluster, and so on. The analysis stops when all objects are classified. Depending on your choice of the cluster radius, you can get too many or too few clusters, or even a single one. By changing cluster radius carefully, several analyses can be performed until the data set becomes tractable.

In quick clustering, the program saves the data for the seed objects in a new file (objects in rows), so that this reduced data set can be subjected to other types of analysis. A label file, containing the original serial numbers, is also saved so that the selected seed objects can easily be identified in these subsequent analyses. Also saved is a group membership vector for all the objects.

> **Note:** For quick clustering the objects should preferably be presented as the rows of the data file, although they can also be arranged as columns (in the latter case, very very large sets may not be analyzed).

5.3.4 Global optimization

The ratio of average within-cluster distances and the average between-cluster distances is minimized during the relocations. Many distance or dissimilarity measures may be used (except the ordinal ones), and the analysis may start from raw data as well as from distance matrices.

If this method is selected from the **Method** menu, then the small box labeled **Number of clusters** is activated near the **Analyze** button, so that the user can enter the desired value.The starting classification may be:

- Randomly generated partition.
- User specified partition (e.g., derived from a previous hierarchical classification using the **Partition from a dendrogram** command in the **Utilities** menu).
- The user provides seed objects, and then each object is linked with the closest seed object to form an initial partition.
- The program selects random seeds.
- The *p* objects that fall farthest apart in the *n*-dimensional space are selected as seed objects for the starting partition for *p* clusters.

Since the final partitions are usually not unique from random starts, you are advised to perform many analyses when the **Number of searches** dialog appears, and then to accept the most optimal result.

5.3.5 Fuzzy clustering

In this case membership is not restricted to one cluster. Instead of assigning a given object into one and only one group, *p* membership weights are calculated for each object, expressing the affinity of the object to all clusters. The membership weights range from 0 to 1.0, and their sum is 1.0 for each object (there is a far analogy with probability).

The number of clusters has to be specified in advance in the small box near the **Analyze** button. Another parameter to be supplied by the user is the *coefficient of fuzziness*: the closer it is to 1.0 the harder (=crisper) the partition obtained. It cannot be 1.0 exactly, because of singularity problems. The method is usually referred to as *c*-means clustering, which minimizes the so-called fuzzy sum of squares of clusters. The membership weights are saved in a new file with objects as rows and groups as columns. The weights imply an ordination of objects with axes as groups, so that the contents of this file can be examined as a scattergram (The **Draw fuzzy cl. as scattergram** button is activated when the analysis is completed). When the number of clusters is three, another option becomes available (**Draw ternary plot**). A ternary plot is a simplex, illustrating the grouping tendency of each object. The tips of the triangle represent the three clusters, and the closer a point to a given tip, the higher its association to the group this tip represents. If an object is positioned in the centroid of the triangle, then it has cluster membership values of 0.3333 with respect to all the three clusters.

5.3.6 Ordinal clustering

The algorithm of ordinal clustering first orders the $m(m-1)/2$ coefficients, so that each d_{jk} is replaced by its rank, r_{jk}. Then, the clustering criterion is examined

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

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. 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 fact, in non-hierarchical clustering 1-*C* is maximized (i.e., *C* minimized) in each iteration step until no improvement is possible. 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*. To do this, first save the dendrogram, and define a partition from this dendrogram (using the **Utilities/Partition from a dendrogram** command). Then, non-hierarchical clustering should be started from this initial partition.

Dissimilarity measures for the ordinal scale type may be used, and the analysis may start from raw data as well as from distance matrices or matrices of ranks of dissimilarities previously saved in **HierClus**.

If this method is selected from the **Method** menu, then the small box labeled **Number of clusters** is activated near the **Analyze** button, so that the user can enter the desired value. The starting classification may be:

- Randomly generated partition.
- User specified partition (e.g., derived from a previous hierarchical classification using the **Partition from a dendrogram** command in the **Utilities** menu).

For the random case, the user has to specify the number of searches. The best result is retained after computations.

5.4 Hierarchical clustering

Whereas most non-hierarchical clustering methods require some *a priori* specification of the number of clusters or other criterion (coefficient of fuzziness, cluster radius), hierarchical classifications do not require such arbitrary decisions. Therefore, these methods are recommended to use first in order to be able to get an insight on the existence of group structure in your data. Typically, hierarchical clustering is represented by a dendrogram, which may be displayed after computations. The dendrogram data are saved in a file for future evaluation or display. SYN-TAX 2000 module **HierClus** performs hierarchical clustering.

5.4.1 Agglomerative methods

These procedures start with *m* clusters, each containing a single object; and then these clusters are amalgamated into larger and larger clusters in each step of the analysis. The fusion of single objects is determined by the distance measure applied, whereas the fusion of clusters is additionally determined by the amalgamation criterion selected.

Distance-optimizing procedures. The cluster amalgamation criterion relates to cluster-to-cluster distances (e.g., single linkage, average linkage, centroid, etc.). The methods are compatible with the Lance-Williams combinatorial equation, so that the raw data are not stored in memory. Alternatively, the analysis may start directly from a distance matrix.

Homogeneity-optimizing procedures. These strategies optimize some internal property of clusters, utilizing either sum of squares, a distance-based or an entropy-based definition. In the latter case only binary data are used (the program automatically transforms "quantitative" scores into binary form). If sum of squares, variance or average within-cluster distances are used, then there is a combinatorial solution so that no raw data are needed during computations, and the analysis may start from a matrix of distances. Information theory clustering, however, requires continuous access to the original data, so it has higher memory requirements.

Global optimization. This strategy takes the ratio of the average of all within-cluster distances and the average of all between-cluster distances. This ratio is in fact a measure of the goodness of the whole classification at various levels. In each step two clusters are amalgamated if their fusion results in the minimum increase of this ratio. The method has a non-hierarchical analogue (Subsection 5.3.4). Raw data or distance matrices are used as input. The dendrogram obtained is similar to the regular dendrograms except that there is no last fusion at the top level: the ratio is undefined for the one-cluster case (i.e., there are no between-cluster distances!). The analysis may be extremely time consuming with large data sets and on slow machines!

> **Note**: In these clustering procedures, there is a built-in procedure to handle *ties* potentially occurring when the minimum dissimilarity is searched in the matrix. Ties may be simply *ignored*, or resolved by *single linkage fusion* of all tied objects, as well as by a *suboptimal fusion*. This ignores the best value and looks for the next most optimal value for which the fusion of objects is unequivocal. The choice among these options is made using a group of radio buttons located on the right side in the **HierClus** module.

5.4.2 Divisive methods

These methods start with a large cluster containing all objects and then divide it into smaller groups as the classification proceeds. The distance of objects is not examined. Instead, information theory measures are applied to express either the "affinity" between pairs of variables or the pooled entropy of variables within each cluster. The division is based on the presence/absence of that variable which has the highest affinity to the others or which produces the greatest decrease of pooled entropy after the division. That is, only binary data are used, and note that the program automatically transforms other data types into binary form.

5.4.3 Minimum spanning trees

A minimum spanning tree is a graph connecting all objects such that there are no circles in the graph and the sum of the lengths of edges is the minimum. In some sense it is also a classification; there is a direct correspondence between single linkage classifications and minimum spanning trees. They are useful, for example, to clarify data structure in 2D ordinations (superposition of trees onto ordinations in **Ordin**). The tree may be displayed to show the topology only, or showing actual branch lengths (use the pop-up menu to switch between these two when the diagram is on the screen).

5.4.4 Additive trees

Neighbor Joining is a cladistic method proposed by Saitou & Nei (1987) to generate optimally additive trees from distance/dissimilarity matrices. In the resulting tree, the sum of lengths of branches along the path between any two objects approximates their input distance. The analysis produces first an unrooted tree, but there are 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.

Rooting is not necessary, however. If you wish, the unrooted additive tree is ouput. 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 additive tree diagram may be displayed after the analysis by pressing the **Draw tree** button. The saved tree file can be used to reproduce the diagram at later time.

5.4.5 Ordinal clustering

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

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

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. 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.

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. To do this, first save the dendrogram and create a partition from it. Then, non-hierarchical clustering should be started from this initial partition.

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*, rather than a weighted dendrogram, being consistent with the ordinality of the previous steps of the analysis.

5.5 Ordination

In general, ordination methods are useful for reducing dimensionality of data structures. The original n variables are replaced by a few artificial variables that explain most of the variation in the data. In this way the relationships among objects are as close to the original as possible, and two- (or three-) dimensional ordination diagrams give a sufficient representation of data structures in many cases. Ordination routines save the scores in new files for future comparisons or display. SYN-TAX 2000 module **Ordin** performs ordination.
The number of axes an ordination method extracts from the data depend on several factors. In most cases, however, the first few dimensions are sufficient to explain most of the variation. The user has the freedom to specify the number of axes to be retained in the **No. of axes** box, near the **Analyze** button.

5.5.1 Metric methods

The metric information contained in the data is preserved by the new configuration. All these procedures are based on the eigenanalysis of symmetric matrices. The new dimensions are linearly uncorrelated and are arranged in order of importance. Although metric methods assume that the original data have a linear structure, the violation of this assumption usually does not influence greatly the interpretability of results.

Principal components analysis (PCA). Raw data are read to compute cross-products, covariances or correlations between variables. After eigenanalysis of this matrix, the eigenvectors are used to find coordinates of objects in a new coordinate system (axes are called the components). Correlations between components and the original variables are also computed and visualized. The ordination of objects, the component correlations and a biplot (in which arrows point to the variables) may be shown by pressing the respective button. In the biplot, the variable scores are rescaled to allow for a clearer display. Note, therefore, that in interpreting biplot diagrams the directions of these arrows rather than the actual positions of their endpoints are important. Object and variable scores are saved in separate files, the save of correlation matrix is optional.

Correspondence analysis (CoA). This method is especially useful to evaluate contingency tables or, in general, to analyze categorical data. A simultaneous ordination of the rows and columns of the data matrix is constructed, thus facilitating the evaluation of relationships between variables and objects. The analysis preserves the so-called chi^2-distances between the rows (and columns). Three weighting options are offered, but the symmetric one is recommended in general.The ordinations for rows and columns, and then the joint plot may be displayed by pressing the appropriate button. Object and variable scores are saved in separate files.

Principal coordinates analysis (metric multidimensional scaling, PCoA). In this case the problem is formulated differently: given a distance matrix of objects a coordinate system is sought in which the original distances are preserved completely such that the first few axes usually provide a fairly good representation of distances. Starting from raw data the distances are to be calculated first. The input matrices should be provided in the same way as for most clustering methods. It is assumed, however, that semimatrices contain dissimilarities or *squared* distances. The coordinates are saved in a new file.

Canonical variates analysis (CVA). The method is also known as multigroup discriminant analysis. In this case there is an *a priori* grouping of objects and the problem is to find linearly uncorrelated axes by maximizing the distinction among groups and minimizing the variance within groups in the new space. There is an option to normalize the eigenvectors such that the resulting group dispersions will be spherical. If the eigenvectors are normalized to unit length, the scatter of objects will be elliptical in the canonical space. The data must be provided such that the variables are columns and objects rows, and the size of group is specified before the first object of each group. The scores are saved in a new file. Each group will appear in a different color when the ordination results are displayed.

Canonical correlation analysis (COR). This method examines the interrelationships between two sets (domains) of variables describing the same objects (e.g., species and environmental variables for sites in ecology). The analysis finds linear combinations for each set of variables in terms of canonical variates such that the correlation (so-called canonical correlation) between the two variates is maximized. The method can be considered as a double principal components analysis followed by the rotation of axes to maximize their correlation. In the data file variables are columns, first the left set variables and then the right set variables.The data should be arranged such that the number of variables on left is equal to or larger than on the right set. A basic requirement is that the number of objects be much larger than the number of variables. The analysis generates scores only for the first two dimensions.

5.5.2 Constrained ordination

Whereas in Canonical Correlation Analysis 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. The first method is recommended for relatively linear data structures (such as those often observed along short gradients) and the second is suitable to long gradients.

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, plus the LC scores. 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.* The two types of variables appear in different colors 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.

5.5.3 Non-metric methods

The metric information in the data or distances is not preserved by the ordination. The method of *non-metric multidimensional scaling* considers the rank order of distances, rather than their actual magnitudes. The objective is to arrange the objects in a few (usually two) dimensions such that the rank order of distances in the new space follows the original rank order as closely as possible. The "goodness" of the non-metric solution is measured by the stress between the original and the new distances. The procedure is iterative, the initial configuration is *random* or *read* from disk file in form of data with objects as rows, in free format. You may start immediately requesting two dimensions; if you choose, say, 3 for start, the two-dimensional solution will be obtained from the three-dimensional. Maximum starting dimensionality is 5. A stress of 0.1-0.2 is considered fairly good, but there is no general rule since the stress is greatly influenced by the number of points. The Shepard diagram gives a visual comparison of original and new distances. In an ideal case the distances increase monotonically over the original distances. The worse the solution, the more scattered the points in the Shepard diagram. After computations, the ordination of objects and the Shepard diagram can be displayed.

Chapter 6

Utilities and related features

Some important possibilities available through the **Utilities** menu were discussed already:

- The options for **Data standardization** are listed in Section 5.1.
- **Export dist. matrix to full format** and **Convert full dist. matrix** are described in Section 3.6.
- The **Transpose data matrix** utility is decribed in Section 3.7.
- The **Excel import/export** utility is discussed in Section 3.8.
- The **Data grid color** command is desrcribed in Section 2.2.

All other commands are described here, although in some cases they were mentioned frequently in this manual (e.g., the Text editor).

6.1 Text Editor

The built-in **Text Editor** is a small Notepad-like, but simpler program that can be evoked from the **Utilities** menu or using the **Edit text file** speed button. An existing data file can be opened and modified in the usual manner, and then printed or saved. Alternatively, a new file can be created if a new name is entered as filename.

6.2 Randomizing raw data

Through the **Utilities** menu , SYN-TAX 2000 offers seven different designs to randomize a raw data matrix. These allow the user to repeat analyses under various randomization designs. These are:

Bootstrap data for rows: In the new matrix, same size as the original, the rows are randomly selected with replacement.

Bootstrap data for columns: In the new matrix, same size as the original, the columns are randomly selected with replacement.

Permute each row: Data values in each row are rearranged randomly in that row.

Permute each column: Data values in each column are rearranged randomly in that column.

Randomize data: All values are randomly relocated in the data matrix (complete randomization).

Random sample of rows: A matrix with fewer rows as the original is created such that the rows are chosen randomly, without replacement.

Random sample of columns: A matrix with fewer columns as the original is created such that the columns are chosen randomly, without replacement.

If you wish, several output matrices are generated, all of them written to the same file and numbered from 1 to n.

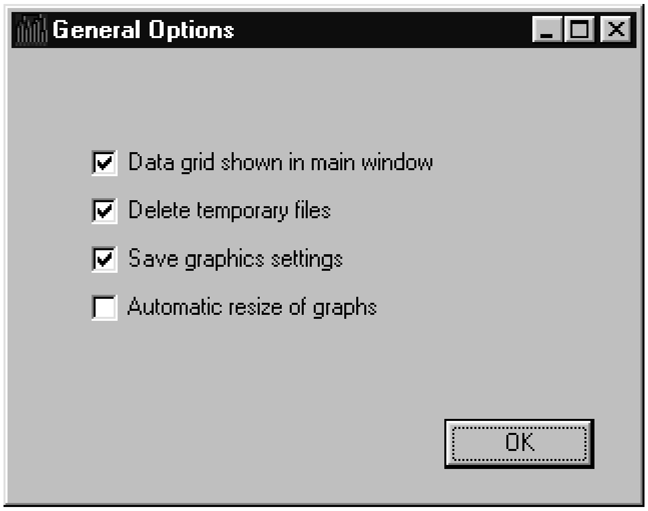
6.3 Partition from a dendrogram

This command, available through the **Utilities** menu in **Ordin** and **NonHier**, allows the user to specify a partition of objects based on a hierarchical classification implied by a dendrogram. After opening the dendrogram file, you enter the number of clusters required in the partition, then the program will "cut" the dendrogram at the appropriate level, starting from the top, of course. The partition is then output in form of a group membership array ('partition file').

For example, if you want a partition into three clusters, the program will cut the tree right below the second highest hierarchical level. In cases with subsequent hierarchical levels being equal, it may happen that unequivocal partitioning is impossible. The user is warned about this.

6.4 General options

In each module of SYN-TAX, there is a possibility to set some general options. The contents of the dialog box is self-explanatory, as shown below on the example of **HierClus**:



☑ Data grid shown in main window

☑ Delete temporary files

☑ Save graphics settings

□ Automatic resize of graphs

6.5 Summary

The **Summary** button in the **Main window** works only if an analysis is completed successfully. After pressing this button, an abbreviated output list of numerical results is typed into a new window. The list is editable so any part of it can be removed. The contents of this window may then be printed on the currently selected printer or saved in a text file.

The default extension of summary files is .TXT.

Appendix A: Filename extensions

For the convenience of the user in distinguishing among many types of files used by SYN-TAX 2000, different DOS extensions are attached to different file types. A full list of default extensions is given below.

Input files

Raw data	*.DAT and *.DTA
Dissimilarity semimatrix	*.DIS
Labels	*.LAB
Partitions (group memberships)	*.PAR
Partition cluster seeds	*.PSD

Output files

Dendrograms	*.DEN
Rooted additive trees	*.ADT
Unrooted additive trees	*.UTR
Minimum spanning trees	*.MST
Ordination score files	*.ORD
Partitions	*.PAR
Summary of numerical results	*.TXT

These default extensions appear first in the filter of **Open** and **Save** dialog boxes. If you wish to use other extensions, then click to the next filter (usually *Text files*, *.TXT) or to the final one (usually *All files*, *.*).

Appendix B: Troubleshooting

Computing errors

Some error messages may appear during computations. In many instances, these errors are due to some of the causes listed below:

1) There is a variable or an object (a row or a column in the data) with all of its values constant (e.g., zero). Remove it from the data. In general, variables with zero variance should be avoided.

2) In RDA or CCoA, the number of constraining (e.g. environmental) variables cannot be larger than the number of criterion variables (e.g., species).

Other errors:

1) The program attempts to read something from Unit 2, then aborts. Without labels, the same program runs normally.

Explanation: The labelfile is bad. If you enter a blank line after the last label, this error does not occur.

2) In the output list, there is "incomplete" information (e.g., percentage of variances in PCA).

Explanation: This is in fact not an error. Many programs print a short output list by default. However, if you wish to see more details, check the **Utilities/General options/Detailed output list** box before the analysis starts.

Graphics errors

Some potentially occurring graphics errors and their explanations are:

1) Text appears horizontally, rather than vertically, in a dendrogram.

Explanation: Specify a **True Type** font for the given graphic item. The program cannot rotate other fonts.

2) After printing a diagram, the title of the diagram appears in a very large font.

Explanation: The printer sent an erroneous message back to your computer. Redraw the figure and the diagram will be normal.

3) No graphics appears. Strange messages, such as "Pointer size must be greater than zero" may appear instead.

Explanation:
- The Ini file of the given application (e.g., Ordin.ini) was modified by an abnormal termination of the program previously, or edited (please do not do that!). Exit from the program, simply delete the respective Ini file and start the program again. In this case, the previous graphics settings are lost, of course.
- You selected zero to be symbol size. This is NOT allowed!

4) In the ordination diagram, full symbols appear even though the empty symbol option was chosen. Also, the symbol color option does not seem to work.

Explanation: This is in fact not an error. In ordinations with grouped objects, only full symbols are used, with their colors automatically set by the program to ensure distinguishability of groups from the background and from each other.

5) Some graphics elements (e.g., arrows in PCA biplots) do not show up.

Explanation: The color of the given graphics element happens to be the same as background color. Use a different color to ensure distinguishability from the background. Alternatively, this offers a good option to hide some graphics element if you do not want to see it.

6) The diagram is larger than the printed area

Solution: The easiest remedy of the problem is that in **Graphics/Line width, printer scaling,** use a printer scaling percentage lower than 100%.

7) There is a "bad" color in the WMF file behind the diagram

Explanation: This color is the panelcolor. The panel color can be modified by the pop-up menu in **Ordin** when a graphics is shown on the screen. Use the same color as for the background, and then your WMF file will be OK. When the diagram is printed, this panel color is ignored.

8) The labels are truncated in graphic displays

Explanation: This is not an error. The number of characters per label is limited to 8 in most cases (trees), the remaining ones are truncated. The exceptions are ordination scattergrams in which labels have the same length as in the input label file. If you find these labels too long, then edit the label file before opening it.

9) Only part of the diagram appears in the graphics window and there is no scroll bar.

Solution: Draw a zooming rectangle backwards (from bottom right to upper left) using the left mouse button.

Some causes of input errors

If the data file is not prepared with caution, the program cannot read the data values appropriately and cannot run. Although some protective checks are built into SYN-TAX 2000, wrong input data may even crash your system under certain circumstances. In most cases, the user receives the ***Error in input file*** message. The following list contains some typical user errors and problems that cause error messages, program aborts or crashes:

1) No title is given in the first row of the data file. It happens when old SYN-TAX data files are used without modifications.

2) No data size is specified in the second row of the data file. Also, it happens when old SYN-TAX data files are used without modifications.

3) There are fewer number of data values than specified. For example, there is no space between some values.

4) Every row of the data matrix must start in a new line in the input file, and this is not the case

5) Non-numbers, e.g. O instead of zero, or capital I or lower case L instead of 1 appear in the data file.

6) Comma was used as value separator, rather than space.

7) The data file was imported from a Macintosh OS, which does not have both *end of line* and *linefeed* characters. In this case, the file must be edited beforehand, i.e., line feed ("enter") added to the end of each input row.

8) Your data file is extremely large and does not fit the available RAM. A potential remedy is to quit all other applications to free some more memory. Also, in case of large data sets much memory can be saved by unchecking the *Data grid shown in the main window* box in the **Utilities/General options** menu.

9) The input file was produced by a previous analysis in SYN-TAX, and due to some error during the computations, non-numbers (NaN-s) or strings of asterisks (*********) appear instead of numbers. For example, ordination score files may occasionally include such strings for high dimensions, which can be resolved by rerunning the analysis with fewer number of axes requested.

Error messages

ACCESS DENIED.
This message occasionally appears if the input file (e.g., *.DAT, or *.LAB) is set to READ Only. By the right button click on the icon of the file, and uncheck the Read Only box in the Properties/Attributes menu.

NOTES