# Multi-Variate Statistical Package

*Version 3.1*

**Users' Manual**

KCS

Kovach Computing Services

# *MVSP Plus*

## Version 3.1

# Users' Manual

Published by Kovach Computing Services, Pentraeth, Wales, U.K.

Printed September 2007

# LIMITED USER LICENCE

Kovach Computing Services ('the licensor') grants the purchaser ('the licensee') a licence for the computer program *MVSP 3.1* ('the program'), in accordance with the terms and conditions contained in this agreement.

The program is licensed for use by one user. The program may be transferred between computers or users, so long as there is NO POSSIBILITY that the program will be used by more than one user at any one time. The licensee may make additional copies of the software for archival purposes only. The accompanying user documentation may not be copied in any way.

The licensee shall not use, copy, rent, lease, sell, modify, decompile, disassemble, otherwise reverse engineer, or transfer the licensed program except as provided in this agreement. Any such unauthorised use shall result in immediate and automatic termination of this licence.

This licence is non-transferrable and non-exclusive. Kovach Computing Services warrants that it is the sole owner of the software and has full power and authority to grant the site licence without the consent of any other parties.

## *Limited Warranty*

Kovach Computing Services warrants the physical diskettes and physical documentation provided under this agreement to be free of defects in materials and workmanship for a period of sixty days from the purchase.

KOVACH COMPUTING SERVICES SPECIFICALLY DISCLAIMS ALL OTHER WARRANTIES OF ANY KIND, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTIES OF MERCHANTABILITY AND/OR FITNESS FOR A PARTICULAR PURPOSE.

The total liability of Kovach Computing Services for any claim or damage arising out of the use of the licensed program or otherwise related to this licence shall be limited to direct damages that shall not exceed the price paid for the program.

IN NO EVENT SHALL THE LICENSOR BE LIABLE TO THE LICENSEE FOR ADDITIONAL DAMAGES, INCLUDING ANY LOST PROFITS, LOST SAVINGS OR OTHER INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OF OR INABILITY TO USE THE LICENSED PROGRAM, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

This agreement does not affect your statutory rights. The agreement shall be interpreted and enforced in accordance with and shall be governed by the laws of England and Wales.

# Contents

## *Suggested Citation*

If you have used *MVSP* in study that you are publishing, the following is a suggested format for the citation:

Kovach, W.L., 2007. *MVSP - A MultiVariate Statistical Package for Windows, ver. 3.1.* Kovach Computing Services, Pentraeth, Wales, U.K.

We are always interested to see how MVSP is being used. We would appreciate receiving reprints of any papers you have published in which MVSP was used for data analysis. Thank you!

## *Address for Correspondence*

Kovach Computing Services
85 Nant-y-Felin
Pentraeth, Anglesey LL75 8UY
Wales, U.K.

E-mail: sales@kovcomp.com
Web: http://www.kovcomp.com/
Tel.(UK): 01248-450414, (Intl.): +44-1248-450414
Fax (UK): 020-8020-0287, (Intl.): +44-20-8020-0287
Note: Please see Chapter 7 before contacting us for technical support.

We also maintain a mailing list for notifying customers of new programs and other items of interest. It is a low volume list, with at most one or two messages a month.

To join the KCS-ANNOUNCE mailing list send an e-mail message to:

listserver@kovcomp.com

with the following text as the subject or first line of the body of the mail message:

subscribe kcs-announce

# Chapter 1 - Getting Started

Welcome to MVSP - A MultiVariate Statistical Package. This is a program for Microsoft Windows™ that performs a variety of ordination and cluster analyses. It provides an inexpensive yet easy means of analyzing your data in fields ranging from ecology and geology to sociology and market research.

MVSP performs several types of eigenanalysis ordinations: principal components (PCA), principal coordinates (PCO), and correspondence/detrended correspondence analyses (CA/DCA). It also does canonical correspondence analysis (CCA), a technique highly popular in ecological studies. It can also perform cluster analysis, using 23 different distance or similarity measures and seven clustering strategies. Diversity indices may be calculated on ecological data; these include Simpson's, Shannon's, and Brillouin's indices.

The number of cases and variables that can be analyzed is limited only by the amount of memory available to Windows (RAM and hard disk swap file), up to a maximum of 2 billion cases and variables.

One possible drawback to ease of use is that some users may be very tempted to take a 'black box' approach to using these statistics, feeding in numbers and coming up with 'The Answer'. We must strongly warn the users of this program that statistics can be DANGEROUS! All these procedures make assumptions about the data and have restrictions on what they can and cannot do. If these assumptions and restrictions are violated, the results could be meaningless. We urge you to become familiar with the methods before you use this program. This manual contains a list of references that we have found very useful in understanding these techniques. In particular, Sneath & Sokal (1973), Gauch (1982), Pielou (1984), Manly (1994) and Davis (1986) are very well written and give very clear discussions of these techniques.

# Installation

To install MVSP 3.1 simply insert the enclosed CD in your CD drive. It should automatically run the main installation program. If it doesn't then simply double click on the "My Computer" icon, double click on the CD drive, then double click the Install program.

You will first be shown a list of available programs on the CD. Ensure there is a tick mark next to MVSP and press the Install button. You will then be asked a few questions, primarily the location to which the program will be installed.

When it is done you can run MVSP by clicking on the MVSP icon in the appropriate Program Manager group or the Start Menu|Programs menu. The first time you run the program you will be greeted by a page of the help file. This will tell you a bit about the program and invite you to follow through a tutorial on how to use MVSP.

### *Uninstalling*

MVSP has an uninstall feature to allow you to remove it. Under Windows 95 and later you can remove it through the Add/Remove programs section of the Control Panel.

# The MVSP user interface

MVSP 3.1 is designed so that everything in a single program window is related to a single data file. When you first load or create a file you will see a single window within the main program frame. This is called the status window and it displays information about the current file. You can then open further windows for various tasks. The combined contents and layout of these windows is called the MVSP Desktop.

## MVSP windows

There are five separate types of windows that can appear within the main MVSP program window. Which windows appear depends on what actions you are performing. The contents of all the windows are related to the currently open data file. When the data file is closed so are all the other open windows.

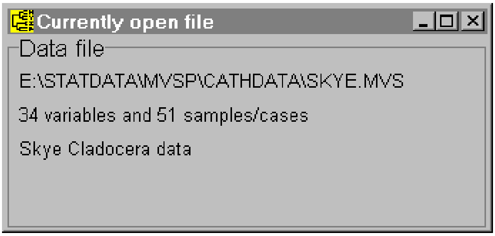
These five window types are Status window, Data editor, Results window, Graphs window and Notepad.

Many of the menu commands will act on the window that is currently active (the one that is topmost among all other visible windows; its caption bar will be colored; all other windows will have grayed caption bars). For example, File|Print will print the contents of the active window and File|Export will export its contents. The Options|Font and Options|Format commands also work on the active window, as do all the Edit menu commands.

Each window maintains its own record of changes to the contents, so that the Edit|Undo and Edit|Redo commands will undo and redo only those changes in the current window.

## Status window

The Status window indicates that a data file has been loaded and gives some information about it. The full name of the file will be given, along with the size of the data matrix and any title that has been assigned to the file.



Currently open file

Data file

E:\STATDATA\MVSP\CATHDATA\SKYE.MVS

34 variables and 51 samples/cases

Skye Cladocera data

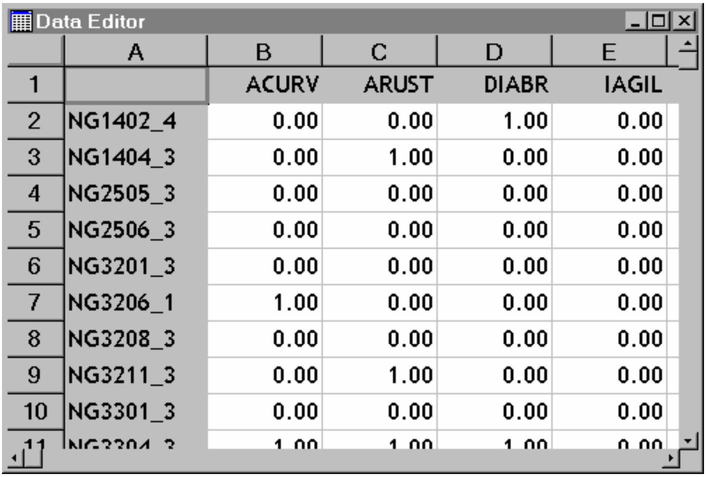
If this window is closed (by choosing Close from the system menu, pressing Ctrl-F4, or (under Windows 95/98 and NT4) by clicking on the close button at the upper left, then the data file will be closed as well.

If a secondary file is open for Canonical Correspondence Analysis (see the section on CCA in Chapter 2) then this window will be expanded so that a second panel appears below the first, with details of that file.

## Data editor

The data editor allows you to view and edit the current data file. it is designed to look and work like a spreadsheet. The variable names are given in the first row (which appears in a different color than the rest of the spreadsheet). Labels for the individual cases can be entered in the first column.

The Data editor is opened by choosing the Data|Edit Data menu item.



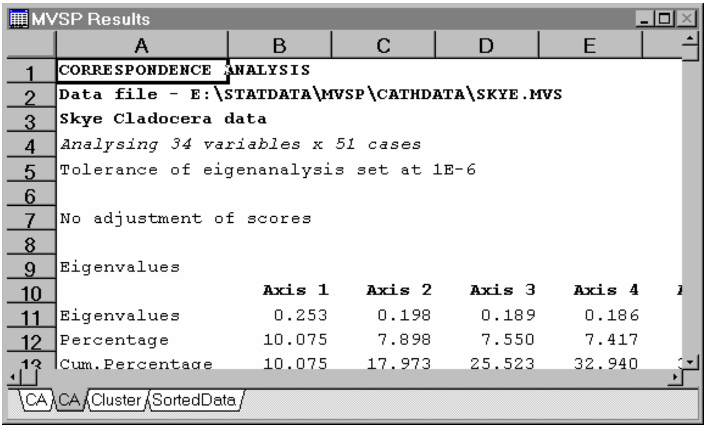
| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | | ACURV | ARUST | DIABR | IAGIL |
| 2 | NG1402_4 | 0.00 | 0.00 | 1.00 | 0.00 |
| 3 | NG1404_3 | 0.00 | 1.00 | 0.00 | 0.00 |
| 4 | NG2505_3 | 0.00 | 0.00 | 0.00 | 0.00 |
| 5 | NG2506_3 | 0.00 | 0.00 | 0.00 | 0.00 |
| 6 | NG3201_3 | 0.00 | 0.00 | 0.00 | 0.00 |
| 7 | NG3206_1 | 1.00 | 0.00 | 0.00 | 0.00 |
| 8 | NG3208_3 | 0.00 | 0.00 | 0.00 | 0.00 |
| 9 | NG3211_3 | 0.00 | 1.00 | 0.00 | 0.00 |
| 10 | NG3301_3 | 0.00 | 0.00 | 0.00 | 0.00 |
| 11 | NG3304_3 | 1.00 | 1.00 | 1.00 | 0.00 |

The actual data are kept in memory as well as displayed in the editor. If the editor is closed any changes you have made will remain in memory (but will not be saved to disk until you explicitly save the data).

If a secondary file is open for Canonical Correspondence Analysis then a second data editor can be opened for editing those data. When two files are open the Data|Edit Data menu item will have a submenu allowing you to specify which file to edit.

## Results window

The results of MVSP analyses are presented in the Results window. It is also works like a spreadsheet, similar to the Data Editor. However, there are two differences. 1) The Results window can have several pages, each containing the results of one analysis. The page to view can be selected by the tabs along the bottom 2) The text and numbers in the spreadsheet cannot be modified, to avoid accidental changes to the results.



```
 1  CORRESPONDENCE ANALYSIS
 2  Data file - E:\STATDATA\MVSP\CATHDATA\SKYE.MVS
 3  Skye Cladocera data
 4  Analysing 34 variables x 51 cases
 5  Tolerance of eigenanalysis set at 1E-6
 6
 7  No adjustment of scores
 8
 9  Eigenvalues
10                      Axis 1    Axis 2    Axis 3    Axis 4
11  Eigenvalues          0.253     0.198     0.189     0.186
12  Percentage          10.075     7.898     7.550     7.417
13  Cum.Percentage      10.075    17.973    25.523    32.940
```

Pages are added automatically when a new analysis is performed. To remove the page that you are viewing choose the Edit|Delete Page menu item (or Delete Page from the context menu that appears when the right mouse button is clicked on the window).

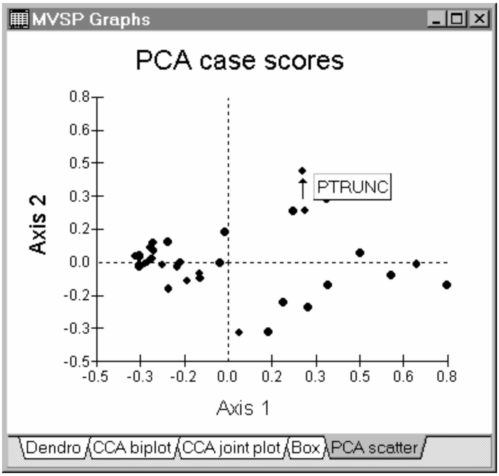
The text on a tab can be changed by double clicking on the tab and editing the text in the resulting dialog box.

The results can be printed (through File|Print) or copied to the clipboard (Edit|Copy) for transfer to another Windows program You can also export the results on the current page to a text file, using File|Export, for importing to other programs.

To save the results for reloading in MVSP later use the desktop saving procedure, described later in this chapter.

## Graphs window

The Graphs window contains the graphs produced from either the results or original data. It is a multipage window, with a single graph on each page. The pages are selected by tabs along the bottom.



The graphs can be extensively modified to suit your requirements. This is done through the Graphs|Edit Graph menu item. This will produce a multi-page dialog box that allows you to adjust many characteristics of the graph.

Changes made to a particular type of graph are saved to be reused next time you produce a similar graph. So, if you modify a scatter plot to have certain sized fonts and types of symbols, the next scatter plot you produce will have those same characteristics.

Pages are added automatically when a new graph is produced. To remove the page that you are viewing choose the Edit|Delete Page menu item (or Delete Page from the context menu.

The text on a tab can be changed by double clicking on the tab and editing the text in the resulting dialog box.

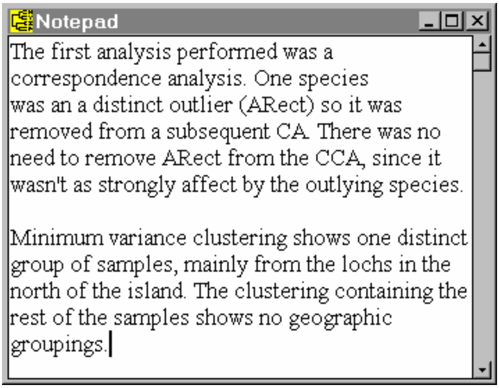
The graphs can be printed (through File|Print) or copied to the clipboard (Edit|Copy) for transfer to another Windows program You can also export the graph on the current page to either a Windows metafile (.WMF) or a .BMP bitmap file, using File|Export, for importing to other programs.

To save the graphs for reloading in MVSP later use the desktop saving procedure.

## Notepad

The Notepad window provides a space for you to make notes about the analyses you've done and any trends you've noticed. It is a simple text editor with word wrap.

You can open and close the window through the Window|Show Notepad menu option. The text in the window is retained even if you close it, so it reappears when you reopen the window (note, however, that it is cleared when a new data file is loaded or created). The text is saved to the desktop file so that it can be reinstated later when you reload the desktop.



The first analysis performed was a correspondence analysis. One species was an a distinct outlier (ARect) so it was removed from a subsequent CA. There was no need to remove ARect from the CCA, since it wasn't as strongly affect by the outlying species.

Minimum variance clustering shows one distinct group of samples, mainly from the lochs in the north of the island. The clustering containing the rest of the samples shows no geographic groupings.

The text in the Notepad can be printed (through File|Print) or copied to the clipboard (Edit|Copy) for transfer to another Windows program You can also export the results to a text file, using File|Export, for importing to other programs.

## MVSP Desktop

The MVSP Desktop is the collection of windows, and their contents, that occur within the MVSP program frame. These include the various pages of results, graphs of the data and results, notes in the Notepad window, and the data editor. The contents of all of these windows are related to the currently loaded data file.

The desktop can be saved at any time, to a file with a .MDK extension. This saves both the contents and positions of the windows. You can then later reload this desktop file and continue your analyses where you left off. You can also have MVSP automatically save the current desktop every time you exit the program and reload it each time you restart, so that you always return to the point where you left off. This is done by setting the “Automatically save and load desktop” option on the Options|Preferences dialog box.

## Creating a new file

To create a new file follow these steps:

1. Select File|New from the menu.
2. Enter the initial number of variables (columns) and cases (rows). More can be easily added later.
3. In the "Data File Type" box make sure that "Regular" is selected.
4. Press OK

A new data file will be created and the Status window (entitled "Currently open file") will be displayed showing that the file is unnamed and giving the size of the matrix.

You may give the file a title, describing what the data represent:

1. Choose Data|Edit title from the menu.
2. Type in the description in the "Main data file title" section of the dialog box.
3. Press OK

The title will then appear in the Status window.

At this point the data will consist of all zeros and the variable and case labels will all be blank. You can now use the data editor to begin entering the data or save the file.

## Entering data

To begin entering data you must open the Data Editor window. Do this by choosing Data|Edit Data from the menu. A spreadsheet-like editor will appear. Initially the cell in the upper left corner will have a dark line around it. This is the active cell and is where any characters or numbers that are typed will appear. You can move the active cell around the spreadsheet, using the mouse to select a new one or by using the arrow keys. Other keystrokes are also recognized for navigating the spreadsheet (see next section).

The spreadsheet has column headers labeled A, B, C, etc. and row headers labeled 1, 2, 3, etc. These headers indicate the position within the spreadsheet and also can be clicked to select the whole row or column for editing purposes. You can change the width of the columns and height of the rows by using the mouse to drag the separator between the headers. Normally each row or column is changed individually. However, holding down the shift key while you resize a row or column causes all others to be changed to the same size (except the first column, which can be different).

The first row and column are shaded and blank when a file is first created. This region is used for entering labels that describe the columns (variables) and rows (cases or samples). These labels are used on the output to identify the variables and cases. You can enter labels by simply making the appropriate cell active and typing any characters. Spaces may be included and the labels may be up to 60 characters long (although shorter labels allow you to keep the columns widths narrow and see more data on the screen).

It is a good idea to enter the labels before the data, particularly the variable names. If you do not enter labels new ones (of the form COL1, COL2, etc.) will be created automatically when you first save the file to disk.

Once labels have been entered you can start typing in the numeric data. This is done by making the appropriate cell active and typing the numbers. You then press the Enter key or one of the arrow keys to finish entering the data; the arrow keys will also move the active cell to a new cell ready to enter the next datum.

You can close the data editor at any time by pressing Ctrl-F4, choosing Close from the system menu, or (under Windows 95/98 and NT4) clicking on the close button in the upper right corner. The data are maintained in memory and will appear when you open the editor again. They are not saved to disk yet, though. You can do this by explicitly saving the data with File|Save Data. If you try to exit the program or open/create a new file you will be warned and asked if you wish to save the data.

## Navigating within the spreadsheet

The data editor spreadsheet provides multiple ways to navigate within the spreadsheet, as specified below.

Some actions depend on whether edit mode is on. Data and labels are edited in place within the spreadsheet cell. With edit mode on, typing new digits or letters will modify the cell contents.

You can enter edit mode by double clicking on a cell with the mouse, by starting to type letters or numbers, or by pressing the either the Enter or F2 keys. Depending on which method you use either a caret (vertical bar) will appear in the cell or the cell contents will be selected.

Edit mode is exited (and the changes saved) by pressing Enter or by moving to a different cell with keystrokes or the mouse. Pressing Esc exits edit mode without saving changes.

| **Key** | **Action** |
|---|---|
| Up Arrow | Moves active cell up one row |
| Down Arrow | Moves active cell down one row |
| Right Arrow | Moves active cell right one column. If in edit mode then moves cursor to the right within the cell. |
| Left Arrow | Moves active cell left one column. If in edit mode then moves cursor to the left within the cell. |

| | |
|---|---|
| Shift+Arrow | Extends selection in direction of arrow key. If in edit mode then selects text within the cell. |
| PgUp | Moves active cell one page up |
| PgDn | Moves active cell one page down |
| Ctrl+PgUp | Moves active cell one page left |
| Ctrl+PgDn | Moves active cell one page right |
| Home | Moves active cell to first cell in row |
| End | Moves active cell to last cell in row that contains data |
| Ctrl+Home | Moves active cell to first row, first column |
| Ctrl+End | Moves active cell to last row and column that contain data |
| Tab | Moves active cell to next cell to the right (or down at end of row) |
| Shift+Tab | Moves active cell to next cell to the left (or up at beginning of row) |
| Shift+Space | Selects current row |
| Ctrl+Space | Selects current column |
| Shift+Ctrl+ Space | Selects entire spreadsheet |
| Shift+Del or Ctrl+X | Cuts current selection or active cell's data to Clipboard |
| Shift+Ins or Ctrl+V | Pastes Clipboard contents into active cell |

| | |
|---|---|
| Ctrl+Ins or<br>Ctrl+C | Copies current selection or active cell's data to Clipboard |
| Enter | If the current cell is not in edit mode then enters edit mode (see above). If cell is in edit mode then exits edit mode, saving changes. |
| Esc | If current cell is in edit mode, abandons any changes and exits edit mode. |
| F2 | Enters edit mode if not already editing. If edit mode is on, cell value is cleared |

## Selecting cells

Several cells in the spreadsheet can be selected at a time for editing purposes. There are a number of ways to do this.

- Use the mouse to point to a cell in the corner of the block of cells to be selected, press and hold the left mouse button, move the mouse to the opposite corner, then release the button.
- Make one of the corner cells active, hold down the shift key and press the arrow keys to move to the opposite corner, releasing the shift key when you are done.
- Single whole rows or columns can be selected by clicking on the row or column header (those labeled A, B, C etc. and 1, 2, 3 etc.).
- Single rows or columns can also be selected by pressing Ctrl-spacebar or Shift-spacebar
- Multiple rows or columns can be selected by clicking on one of the headers, then holding down

the mouse button and moving the cursor to another header.

- The entire spreadsheet can be selected by clicking on the button in the upper left corner of the spreadsheet or by pressing Ctrl-Shift-spacebar.

## Editing data

There are a number of editing actions you can perform on existing data.

### *Overwriting old data with new*

Make the cell active (click with the mouse or move to that cell with the cursor keys), type the new value, and press Enter.

### *Modifying data within a cell*

Make the cell active and press F2 (or double click on the cell with the mouse). This will cause the contents of the cell to be selected. Clicking again or pressing the right or left arrow keys will cause a caret (vertical line) to appear. You can move this back and forth with the cursor keys.

Any new letters or numbers typed will appear at the position of the cursor. With labels, and the portion of numeric data to the left of the decimal, the new characters will be inserted. With the fractional part of numeric data the typed digits will replace those at the caret (which will be shown as a triangle under the digit).

### *Clearing data in one or more cells*

Make the cell active and press the Del key, choose Edit|Clear from the menu, or choose Clear from the context menu. All cells must have some content, so with numeric data the cell value will revert to a 0.00. Labels revert to Case or Var.

To clear several cells at once select the desired cells and use one of the above commands.

### *Pasting data from the clipboard*

Make active the cell at the upper left of where you wish the clipboard contents to be pasted. Then choose Edit|Paste, press Ctrl-V, or choose Paste from the context menu. The pasted data will overwrite existing data.

Any non-numeric data pasted into a numeric cell will be ignored; the cell will display 0.00 instead. If the data in the clipboard contains more rows or columns than will fit in the spreadsheet then the remaining data in the clipboard will be ignored. You should first add new rows or columns if needed before pasting.

### *Adding new columns and rows*

New rows or columns can be added singly by going to the last cell in the row or column and pressing the left or down arrow keys.

To insert new rows or columns in the middle of the spreadsheet select a whole row or column, then choose Edit|Insert Row/Column (or Insert Row/Column from the context menu). To insert two or more rows or columns select two or more existing ones.

If Insert Row/Column is chosen when there are not any whole rows or columns selected a dialog box will appear. This lets you specify exactly what to insert.

### *Deleting columns and rows*

To completely remove rows or columns from the spreadsheet select a whole row or column, then choose Edit|Delete Row/Column (or Delete Row/Column from the context menu). To delete two or more rows or columns select two or more existing ones. **(Note** - pressing the Del key while columns or rows are selected will not remove them, but instead will delete their contents).

If Delete Row/Column is chosen when there are not any whole rows or columns selected a dialog box will appear. This lets you specify exactly what to delete.

### *Reversing editing changes*

The MVSP data editor keeps track of all changes you have made. Changes can be undone, in the reverse order that they were performed, by choosing Edit|Undo from the menu (or pressing Ctrl-Z). The menu item and the hint on the status bar at the bottom of the window will tell you exactly what change will be undone next.

To reinstate a change that you have undone choose the Edit|Redo menu item.

The list of editing changes is only maintained while the data editor is open. If you close the data editor and reopen it you will not be able to undo previous changes.

## Saving data

Data are saved through the File|Save Data menu item. If you have created a new data file (and the status window says "Unnamed" for the filename) you will then be presented with a standard Windows Save As dialog box. You can choose the appropriate disk and directory, then type a name into the "File name" edit box and press OK.

If the file already has a name (which will appear in the status window) then choosing File|Save Data will save any changes directly, with no dialog box.

If no changes have been made to the data the Save Data menu item will be dimmed and cannot be selected. You can tell if the data have been modified by looking at the status bar when either the status window or data editor are the topmost window; the word "Modified" will appear in the second panel from the left. Note that if the results, graphs or notepad windows are topmost then this modified indicator will reflect their modified status, not that of the data.

To save the data under a new name choose File|Save As, which will produce the dialog box described above. Make sure that the "Save file as type" section of the dialog is set to "Data files (*.MVS)"

Note that if a second file of environmental data (a *.MVE file) is open for use in CCA then File|Save Data will save both files. It will ask for files names if either or both files are still unnamed. File|Save As will normally only allow you to save the regular data file (*.MVS) under a new name. To save the environmental data under a new name you must first make sure the environmental data editor is the active window before choosing File|Save As.

## Loading an existing file

An existing file can be reloaded from disk through the File|Open command. This will display the standard Windows File Open dialog box. You can choose the appropriate disk and directory, then select a file from the list of files or type a name into the “File name” edit box and press OK.

By default the list will include all MVSP files. This includes regular data files (*.MVS), CCA environmental files (*.MVE), symmetrical matrix files (*.MVD) and desktop files (*.MDK). You can restrict it to showing only one type of file through the “List files of type” section of the dialog.

When you load a new data or desktop file the currently loaded file (if any) will be closed. If you select to open a MVE file while another data file is already loaded it will attempt to open it as a secondary file; this will be interpreted as being the environmental data for use in canonical correspondence analysis. If you try to open a MVE file when there is no other file open then it will be treated as a regular data file and can be used in any analysis.

You can also open new files by drag and drop. You can drag files from the Program Manager or Windows Explorer onto the MVSP program icon or onto the window of an already running MVSP program. You can also double click a MVSP file in the Program Manager or Windows Explorer to start MVSP and load the file.

## Performing an analysis

Once a data file has been created or loaded you can then perform analyses. The Analysis menu lists the various types of analyses available. Selecting any one of these will cause a dialog box to be displayed, giving the various options for the analysis. When the OK button is pressed a new page will be added to the Results window and the analysis will be performed.

Here is an example of the steps required to perform a principal components analysis:

1. Open a data file (the FOOD.MVS example data file distributed with MVSP can be used).
2. Select Analyses|Principal Components Analysis from the menu. A multi-page dialog box, with tabs along the top, will appear.
3. To discover what each option does you can use one of two help systems. Clicking on the ? button in the upper right corner of the dialog box, followed by clicking on one of the options on the dialog, will produce a small window with an explanation of that option. Alternatively, pressing the button labeled "Help" will display a help page describing all the options.
4. On the options page set the appropriate options. For example, we would usually want "Center data" option to be ticked. We may want to use "Data transformation" to log transform the data before analysis.
5. You usually will not need to adjust any options on the Advanced page, but you may wish to go to that page to select the "Transformed data" option, which will result in the log transformed data being printed out along with the results.
6. By default all cases and variables in the file are included in the analysis. If you wish to drop some of the variables or cases go to the Select page of the dialog. All the variables and cases will be listed, with a tick mark before each. To

drop a variable just click the box containing the tick mark so that it is cleared.

7. Press OK. The Results window will open, if it isn't already visible. A new page will be added, to which the results will be written.
8. As the analysis proceeds messages will be displayed on the status bar indicating the current stage of the analysis. The progress bar, also on the status bar, will indicate the percentage of the analysis finished.
9. If you wish to cancel an analysis before it is finished you can press the Esc key or click on the cancel button next to the progress bar.
10. Once the analysis is finished you can look at and print the results, export them to a file for inclusion in another program, or you can graph them.

## Producing a results graph

Diagrams are produced through the Graphs menu. It has three basic types of graphs. We will produce a scatter plot and a scree plot of our PCA results.

1. Perform a PCA analysis.
2. Make sure the page of results we want to graph is visible. Also, if the data editor is open make sure it is not the active window. If the data editor is active then the scatter plot command will produce a graph of the original data rather than the results.
3. Choose the Graphs|Scatter Plot menu item. A dialog box will appear.
4. Use the drop-down boxes in the "Axes to plot" section to select the axes to use for the graph. By default the first PCA axis will be along the bottom of the graph (the X axis) and second PCA axis will be on the side (the Y axis). A three-dimensional scatter plot can be produced by setting the Z axis to a third PCA axis.

5. Set the "Plot type" to "Cases" to get a plot of the cases.
6. Press OK. The Graphs window will open, if it isn't already visible. A new page will be added, to which the plot will be drawn.
7. Click on one of the points on the scatter plot. A small box will appear giving the case label for that data point. Click on a few other points. To turn off the label box just click on the background of the graph.

### *Customizing*

8. Let's customize the appearance of the graph. Choose the Graphs|Edit Graph menu item. A multi-page dialog box will appear.
9. Go to the Fonts page by clicking on the tab at the top labeled "Fonts". Set the "Apply To" section of the dialog to "Graph Title", then click on the left end of the scroll bar in the "Size" section to reduce the size of the title font. Set "Apply To" to "Other titles" and reduce the size of that as well, then do the same for the "Labels" font.
10. Go to the "Markers" page. Use the "Symbols" drop-down box to choose a filled circle to use for points on the graph. You can use the "Size" and "Color" options to modify those characteristics as well.
11. Go to the "Labels" page. In the section labeled "Data Labels" click the "On" box so that it is ticked. This will cause the data points to be labeled with the case labels in the original FOOD.MVS file.
12. Go to the "Axis" page. Select the "X Axis" option under "Apply to Axis", then select the "Bottom" option under "Position". Go back to the "Apply to Axis" section and choose "Y Primary" and select "Left" under "Position". This will put the axes and scales to the bottom

and right of the graph, rather than forming a cross through the middle.

13. Click OK. The graph will be redrawn with the new options.
14. Produce a second scatter plot by following steps 3-6 again. Note that the changes to the graph appearance you just made will be used for this new graph as well.

### *A second plot*

15. Now let's do a scree plot. Choose the Graphs|Scree Plot option. A graph will be drawn immediately, with no dialog box (this type of graph has no options).
16. Note that this type of plot is most informative if all axes have been extracted in the PCA, by setting the "Axes to extract" option of the PCA dialog to "All".

## Producing a data graph

Graphs of the original data can also be produced, allowing you to examine individual variables.

1. Open the data editor by selecting Data|Edit Data from the menu. Make sure the editor is the active window.
2. Select Graphs|Box and Whisker Plot. A dialog box with a list of all the variables will appear.
3. If you wish to restrict the graph to just a few variables then click on the boxes to the left of the ones you don't want, so that the tick mark is cleared. Otherwise leave them all ticked.
4. Press OK. A box and whisker plot will be drawn.
5. Follow steps 8 and 9 under "Customizing" in the preceding section on "Producing a Results Graph" to customize the fonts on this graph. The "Labels" font will adjust the labels at the bottom of this graph.

6. Go to the "Markers" page of the graph editing dialog and change the "Color" of the boxes.
7. Go to the "Style" page and tick the "Show Samples" checkbox. This will place points on the boxes to indicate the positions of the data points. You may also want to adjust the "Symbol" and "Symbol Size" on the "Markers" page so that they don't obscure the boxes too much.
8. Press OK. The box plot will be redrawn.
9. Now let's produce a scatter plot of two of the variables. First make sure the data editor is the active page.
10. Choose Graphs|Scatter Plot from the menu.
11. In the resulting dialog box use the drop down boxes to set the variables to be plotted on the X and Y axes; let's choose Garlic and Oil from the FOOD.MVS file.
12. Press OK. The graph will be drawn.
13. Go through steps in the "Customizing" portion of the preceding section on "Producing a Results Graph" to do similar customizations.

## Printing

Now let's print the results so that we can study them in detail.

1. Make the Results window the active window and ensure that the page you want to print is visible (using the tabs along the bottom).
2. Select File|Print from the menu. The standard Windows print dialog will be shown. Its appearance will differ depending on the version of Windows you are running.
3. Let's assume your default printer is a monochrome laser printer. Make sure that the print dialog has this printer selected.
4. Press OK. The results will be printed.

5. Now let's print a graph. Make the graph window active and select the graph you wish to print.
6. Select File|Print from the menu.
7. Now let's say you have access to a color printer across a network. Let's also assume you have Windows properly set up to use this printer. It should appear on the list of printers in the standard Windows print dialog. Select that printer and press OK. The graph will be printed.
8. Note that the printer used for each type of window is saved, so next time you wish to print a graph it will by default use the color printer you selected.

## Saving your work

It's time to quit for the day, so let's save our work so far. We have two options. We may not be interested in saving the results and graphs we've produced, since we have now printed them out. In this case just follow the procedure under the section Saving Data earlier in this chapter.

Alternatively we may want to save everything, so we can go back to modify and print the graphs for publication. This is done through the Desktop saving procedure, which saves the position and contents of all windows currently visible in MVSP. To do this:

1. Choose File|Save Desktop from the menu.
2. If we had previously saved or loaded the desktop then it will already have a filename, so the desktop will simply be saved to that file with no intervention. (Note - the name of the desktop file will appear on the caption bar at the top of the main MVSP window).
3. If we have not yet saved the desktop then we will be prompted, with the standard Windows

Save As dialog box, to give it a name. Type a name in the "File name" section.

4. Press OK. The file will be saved.

We may want MVSP to automatically save our desktop each time we exit the program, so that we can pick up where we left off each time we run MVSP. To do this:

1. Choose the Options|Preferences menu item.
2. In the resulting dialog box click the "Automatically save & restore desktop" option so that it is ticked.
3. Press OK. Now when we exit MVSP it will save all our work. If we have previously saved the desktop, so that it has a name, then everything will be saved with no intervention. Otherwise we will be prompted for a name.
4. Restart MVSP. The desktop file will be reloaded and all the windows will reappear just where they were before.

## Reloading your work

If you do not have the "Automatically save and restore desktop" option selected then you can use the manual method of reloading the desktop. Simply:

1. Choose File|Open from the menu.
2. Use the Drive and Directory sections of the resulting standard Windows Open File dialog to locate your files, then select the appropriate .MDK file from the list. You can restrict the list to just desktop files by changing the "List files of type" section to "Desktop files".
3. Click OK. The file will be loaded and your desktop restored.

## Copying to other Windows programs

Let's say you now want to transfer your results and graphs to your word processor for inclusion in a report.

The simplest way is by copying to and pasting from the Windows clipboard.

1. Start up your word processor and load (or create) a file for your report.
2. Go to the MVSP Results window and turn to the page you want to transfer.
3. If you want to transfer the entire page to your word processor then choose the Edit|Select All menu command. Otherwise use the mouse or keyboard to select the portion of the results you want to copy.
4. Choose the Edit|Copy menu command.
5. Switch to your word processor, place the caret at the point where you wish to insert the results and choose Edit|Paste. The text will be inserted into your document.
6. The text forming the columns of the results spreadsheet will be separated by tabs. How these are interpreted and shown in your word processor will depend on the program and how you have the tab stops set up. You may need to do some work to make the results look neat. For example, in Microsoft Word™ you can use the Table|Convert Text to Table to make the results into a neat table. If you are pasting into a spreadsheet program rather than a word processor then the columns of the MVSP window will automatically go into columns of the spreadsheet.
7. Now let's transfer a graph. Switch back to MVSP and bring the Graphs window to the top, choosing the correct page for the graph you want.
8. Choose the Edit|Copy menu item.
9. Go back to the word processor, go to the place where you want to insert the graph, and choose Edit|Paste. The graph will now be inserted into the document.

Instead of copying directly you may wish to save the results and graphs to a file in a form where they can imported to other programs later. To do that follow the procedure in the next section.

## Exporting results and graphs

Once you've created some results and graphs you may wish to save them in a form that can be used by other programs. You may want to include the results in a report, or load a graph in a drawing program to do some further customizations. This is done by exporting. Let's export one results page and one graph.

1. Make the MVSP Results window active and turn to the page you want to export.
2. Choose the File|Export menu command.
3. Select the drive and directory where you wish to save the results, then type a name into the "File name" box.
4. Press OK. The file will then be created and the results exported.
5. In the exported file the text forming the columns of the Results spreadsheet will be separated by tabs. Most spreadsheets will automatically separate the numbers into columns when importing these files.
6. Now let's export a graph. Bring the Graphs window to the top, choosing the correct page for the graph you want.
7. Choose the File|Export menu item.
8. Select the drive and directory where you wish to save the results.
9. Set the type of output file through the "Save file as type" section. The Windows Metafile option save the graphs as lines, symbols and text, so that it can be easily modified and rescaled in a drawing program. The Windows Bitmap option saves an exact picture of what is on the screen, made up of a series of pixels. The

elements of a bitmap cannot be as easily modified as a metafile, although you can use a painting program to add new things to the graph or erase some parts.

10. Type a name into the “File name” box.
11. Press OK. The graph will be exported.

The contents of the Notepad window can also be exported in a similar way, to a plain text file.

## Importing data

MVSP allows you to import data from a number of sources. The Import Preview dialog box lets you see how the data will be imported and to change various options, making importing data trouble-free. Let’s try importing data from a spreadsheet:

1. Choose the File|Import menu item. A standard Windows Open File dialog box will be displayed.
2. Go to the “List files of type” section and select the type of file to import. Several spreadsheet and database file types are listed, along with plain text and Cornell Ecology program files. We will select the Excel file type.
3. When the file type is selected all the files of that type in the current directory will be displayed. Find the file you wish to import and click on it so that its name is displayed in the “File name” box.
4. Press OK. You may first get some warning messages; just click on OK for these. MVSP will open the file and display the Import Preview dialog box. The spreadsheet-like section at the bottom of the dialog will show how MVSP initially thinks it should import the data.
5. If everything looks OK (i.e. the numeric data seems to be displayed properly and the appropriate variable and case labels are displayed in the shaded parts at the top and

left) then simply press OK. The data will be imported and the Status window will be displayed.

6. If the formatting of the numeric data looks odd (for example what you expect to be 0.16 might be displayed a 0.159999999999) try ticking or unticking the "Import as strings" option.
7. If your spreadsheet does not have labels for the cases, and it is trying to use your first column of data as labels, then tick the "Create new row names" option. The preview will be adjusted and new labels, of the form Row1, Row2, etc., will be created.
8. If your column/variable names are not appearing in the shaded part at the top of the preview then adjust the various "Column names" options until it can find the appropriate labels. If your spreadsheet does not have any labels then you can use the "Make up names" option.
9. If the data being imported have a grouping variable (see section on "Working with grouped data" below) then tick the box entitled "Grouping variable, column:". You must then set the number specifying which column contains the grouping variable. This column will be shaded in the preview grid.
10. The spreadsheet file may be a multipage file, with the data you are interested in on the third or fourth page. To use the correct page go to the "Spreadsheet Options" page of the dialog and use the "Sheet" drop-down box to select the correct page. You can also use this section to specify a particular range of data to import.
11. Press OK. The data will be imported and the Status window will be displayed.

Note that if MVSP encounters non-numeric data in a part of the spreadsheet that it believes should contain data it will give a warning message. The preview will display the actual text found in the file, but when it is imported any non-numeric spreadsheet cells will be converted to 0.

Now let's try importing a text file:

1. Repeat steps 1-5 above, using the Text file type instead of Excel.
2. If the numeric data do not appear properly in the preview section (for example, the cells may all be blank or contain more than one value) then go to the "Text Options" page of the dialog. Here you can change the field delimiter to another type to match that used in the text file.
3. If the data are still not displayed properly you may also need to change the "Quote character" (the character used to delimit text labels that have spaces in them; usually a double quote) or the "Decimal place" character (by default a period but could also be a comma).
4. Repeat steps 7 and 8. The "Column names" options will be slightly different but will act in a similar manner.
5. Press OK. The file will be imported.

## Exporting data

Data can be exported to a variety of file types, including spreadsheets, databases and various types of text files. You can also export to the MVSP version 2 format used by the DOS version of MVSP.

To export data:

1. Ensure that the Status window or the data editor is active (if the Results, Graphs or Notepad window is active then the contents of that will be exported instead).
2. Choose the File|Export menu item.
3. In the resulting dialog box set the "Save file as type" to the type of file you wish to export.

4. Locate the correct drive and directory, if necessary, and type a name into the "File name" box.
5. Press OK. The new file will be created.

Note that the files created will be of the earliest file version for that type. For example, the Excel files created are Excel 2.1, rather than the more recent Excel 5 or Excel 97. This provides the widest possible compatibility, since these programs can easily read earlier versions of their files. The other versions created are: Lotus v1, DBase III, Paradox v3.0, Quattro v1.0, and Symphony v1.0

## Working with grouped data

In many situations, cases in a data matrix can be assigned to groups. Examples included samples from three geographic areas, specimens of five different species, or measurements taken in ten different years. MVSP now allows you to set up groups and assign each case in a data file to one of these groups. These groups are then displayed on the output and can be plotted separately on resulting graphs.

### *Adding a grouping variable to a data file*

Groups in a data file are identified by a separate grouping variable. When creating a new file you can specify that the file should include groups by ticking the "Include grouping variable" in the New File dialog box.

A grouping variable can also be added to an existing data set. When the data editor is open and active, choose the Data|Add Groups menu option.

Note that a grouping variable can not be added to a CCA environmental data file.

### *Setting up groups*

Once groups have been added to a file there will be two changes to the data editor. First, a new, shaded column will appear in the editor, just to the right of the case labels. Also, a Data Groups "palette" (a small floating window) will appear on the screen.

To set up groups you must first enter the group names into the palette. These names then appear in the drop-down boxes in the Groups column in the data editor. This means that you do not need to retype the group name for each case. You also don't run the risk of accidentally creating a "new group" through a typographic error.

To set up groups follow these steps:

1. If you have just added a grouping variable, the palette will contain a group called "First Group". You can give this the name of your first group by clicking on it, typing a new label, then pressing Enter.
2. A second group can be added by pressing the add button (marked with a "+" symbol). This adds a new row to the palette, where you can type the new group label.
3. Continue adding groups until all have been added. You can add more groups later if necessary.

### *Specifying group for each case*

Once you have set up a few groups you can assign a group to each case in the data file. There are two ways of doing this.

Each cell in the Groups column of the data editor has a downward pointing triangle to the right. If you click on this a list of the current groups is displayed in a drop-down box. To make the assignment simply click on the group to which the current case belongs. That group name will then appear in the Groups column.

You can also assign cases to groups using the keyboard. When the desired cell in the Groups column is selected press the Enter key. This will cause the drop-down box to be displayed. You can then use the arrow keys to move up and down through the list to select the correct group. Pressing Enter again will assign that case to that group.

If you have a long list of groups you can quickly go to a particular group name by typing the first letter of the group. If several group names have the same first letter you can cycle through them by repeatedly pressing that letter. If most or all of your groups have different first letters then data entry can be speeded up by simply typing that first letter when the cursor is in the Groups column.

### *Changing group labels*

You may wish to change the name of one or more groups, either to describe them better or to correct spelling mistakes. This can be done through the Data Groups Palette. Simply click twice on the group that you wish to rename. This will put the palette into edit mode, where you can then type a new name or use the arrow and backspace keys to edit it. Pressing Enter will finish the editing. At this point that group name in the Groups column of the data editor will also be changed to reflect the new one.

### *Deleting groups*

Groups can be deleted individually by clicking on the name you wish to delete, then clicking on the Delete button (marked with a "-" symbol). When this is done all the cases in the data editor that were assigned to that group will now be unassigned. This is indicated by the cell in the Groups column being blank.

### *Moving and hiding group palette*

The group palette can be moved anywhere on the screen (even outside the MVSP window) so that you can easily view it while editing data. Its position will be remembered the next time you run MVSP.

The group palette is also a "roll-up" palette. It can be reduced so that just the title bar (which is smaller than most window title bars) is visible. This is done by clicking on the button with the upwards pointing arrow (↑) on the title bar. The window can be displayed again by clicking the down arrow button (↓).

If you wish to have the group palette completely disappear then use the Data|Show Group Palette menu command. When the palette is visible this will be ticked. Click it again to remove the tick and hide the palette.

### *Removing grouping variable*

If you decide that you do not wish to have a grouping variable with this data file you can remove it using the Data|Remove Groups menu item.

The grouping variable will also be removed automatically if you transpose the data matrix (since it is unlikely that the groups for the rows will exactly correspond to the columns). If you wish to have groups in the newly transposed matrix you can add them with the Data Add Groups menu item.

## Displaying groups on graphs

Graphing results from grouped data works in the same way as for regular data. For scatterplots, the groups defined in the data file are automatically displayed using different symbols. In dendrograms, the case labels at the end of each branch have the appropriate group name appended.

### *Modifying symbols for each group*

By default, the symbols for each group are different shapes and colors. The shapes and colors used can be modified through the Graphs|Edit Graph menu item. When the Edit Graph dialog is displayed, select the Markers page. An example scatterplot, showing the different symbols used, will be displayed. Click on the symbol you wish to change; the Set number displayed in the “Apply to” box below the graph will change to show which set is selected. Then, change the color, shape and size using the controls to the right of the graph.

### *Modifying the group legend*

When groups are being plotted a legend will be displayed showing which symbol goes with which group. This legend can be modified and moved to different positions through the Edit Graph dialog box.

First, choose the Background tab of the Edit Graph dialog box. In the lower left is a section entitled "Legend". You can use the buttons there to position the legend at any side of the graph or in any corner. The scroll bar labeled "Size" lets you alter the amount of area covered by the legend.

Also on this dialog box page, you can change the appearance of the legend. First select "Legend" in the "Apply to" section, then use the controls in the "Style" section to modify the type of box drawn around the legend.

The typeface and size of the font for the legend can be changed in the "Font" page of the dialog. Again, first select the "Legend" option in the "Apply to" section, then modify the font settings as appropriate.

If you wish to eliminate the legend from the graph you can do this through the Data page. Click the "Legend Text" button; this will display the group names. Delete the text of the first group name (that in the leftmost box), then click OK. The entire legend will then be removed.

## Working with symmetrical matrices

Symmetrical matrices are the distance or similarity matrices required for clustering and PCO. Normally MVSP 3 calculates these automatically and discards them after the analysis is finished. This is unlike MVSP 2, where you had to first run the distances of similarity procedure, save the symmetrical matrix to disk (with a

*.MVD extension), then load this to do clustering or PCO. MVSP 3 can load these old MVSP 2 *.MVD files, as well as create its own. This is useful if you are calculating distances or similarities with another program and wish to analyze them with MVSP's clustering or PCO. See the section on importing symmetrical matrices, later in this chapter, for tips on transferring distances or similarities into MVSP.

You can use File|Open to open .MVD files. If you currently have another file open (raw data or another MVD file) it will be closed. When a MVD file is open some menu options will not be available as they are not applicable to this data type. These included some analyses and most of the data manipulation routines. Also, some of the options on the clustering and PCO dialog boxes will be disabled.

If you open the Data Editor window you can then view and even modify the symmetrical matrix. The diagonal elements of the symmetrical matrix will be highlighted by a different colored background. The data editor enforces the symmetrical nature of the matrix, so that if you change a data value below the diagonal the corresponding one above the diagonal will also be changed. If you change one of the diagonal elements all the others will be changed to match it. If the diagonal elements are 0.0 then it is assumed that the matrix is a distance matrix. Any other value means that the matrix is a similarity matrix and the value in the diagonal is the maximum similarity.

Once you have made any changes you can save the matrix to a MVD file in the usual way (File|Save Data or File|Save As). You can also have MVSP 3 save the distances or similarities calculated during an analysis by ticking the “Save Dist./Simil.” option on the analysis dialog box. The type of matrix to be used for saving (e.g. upper half matrix, lower half matrix with diagonal, etc.) can be set through Options|Preferences.

## Symmetrical matrices

Symmetrical matrices record distances, similarities or other measures of association between various cases or variables, in a pairwise manner. They have the same number of rows as columns, with each case (or variable) being represented by one row and one column.

The distance or similarity between two cases is recorded where the row for one of them intersects with the column of the other. Thus the distance between the first and second cases in the matrix will be in the second row of the first column. The same distance will also be in the second column of the first row; this symmetry is where this matrix type gets its name.

The diagonal of a symmetrical matrix is where the row and column for a particular case intersects with itself. For distance matrices the values along the diagonal will be 0, since by definition there is no distance between an object and itself. For similarity matrices the value along the diagonal will be the maximum similarity, usually 1 or 100.

A symmetrical matrix can be fully specified, with both symmetrical halves being included:

***Full matrix***

```
0     1     1     1     1
1     0     2     2     2
1     2     0     3     3
1     2     3     0     4
1     2     3     4     0
```

However, often only one half of the matrix is given, since there is no need for repeating the distances twice. The half presented can be either from above or below the diagonal. Also, the diagonal does not necessarily need to be included, although in this case you must make an assumption as to the value of the diagonal. The four possible combinations of half matrices are given below.

### *Lower half with diagonal*

```
0
1     0
1     2     0
1     2     3     0
1     2     3     4     0
```

### *Lower half without diagonal*

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

### *Upper half with diagonal*

```
0     1     1     1     1
      0     2     2     2
            0     3     3
                  0     4
                        0
```

### *Upper half without diagonal*

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

You may select which type of matrix MVSP uses to save .MVD files through the Options|Preferences dialog box.

## Importing symmetrical matrices

Currently MVSP 3 cannot import symmetrical matrices with File|Import. All files imported in this way are assumed to be rectangular matrices.

You can, however, use the Windows Clipboard to transfer data from other programs:

1. Select the File|New menu item.
2. On the resulting dialog box select the Symmetrical Data File Type and set the number of variables to the appropriate number.
3. Press OK.
4. Open the data editor with Data|Edit Data.

5. Switch to your spreadsheet or other program containing the symmetrical matrix.
6. Select the entire numeric part of the matrix; it will probably be easier to transfer the labels separately.
7. Choose the Edit|Copy menu item,
8. Switch back to MVSP and make sure the data editor is the active window.
9. Place the cursor in the appropriate cell of the data editor, depending on the type of symmetrical matrix you are pasting. If the diagonal is present then place the cursor in cell B2. If it is not then place it in cell B3 if it is a lower half matrix or C2 for a upper half matrix.
10. Choose the Edit|Paste menu item to paste the data.
11. Go back to the spreadsheet or other program, select the labels and choose Edit|Copy.
12. Return to the MVSP data editor and paste the labels, starting at either Column A/Row 2 or Row 1/Column B, depending on their orientation in the original program.
13. Choose File|Save Data and type in a name for saving to disk. The .MVD extension will be used automatically.

If the symmetrical matrix you wish to import is instead in a text file you can modify it with a text editor so that it is in the correct form to be read by MVSP. See the section on symmetrical matrices in Chapter 6 for more information.

## Performing Canonical Correspondence Analysis

Canonical Correspondence Analysis (CCA) is a special ordination method, widely used in ecological studies, that performs a constrained ordination using two data matrices, one of species occurring in various samples and one of the environmental characteristics of those samples. This results in an ordination in which the species axes are constrained to be linear combinations of the environmental data. In this way, the relationships between the species distributions and the environmental parameters are an integral part of the ordination, rather than being interpreted afterwards as in most ordination techniques.

The requirement for having two data matrices means that preparing to do a CCA is a bit different than for other analyses. In particular, two data files must be open and the regular data must be distinguished from the environmental data.

Files with an extension of .MVS are the standard MVSP data files. Those with a .MVE extension are special files for use with CCA. MVE files have the same structure as MVS files; in fact, you could simply rename the MVE files as MVS and treat it as a normal data file. The different extension is so that, when you try to load a MVE file, the program knows you wish to load a second matrix. Normally when you try to load a file it will first close the existing one.

The easiest way to create a new environmental data matrix is to first open or create a regular data matrix. In this way MVSP can help by giving the new environmental data matrix the same number of samples and the same sample labels (CCA requires that both matrices have the same number of samples and that they are in the same order). The steps for doing this are:

1. Load (or create) a MVS file.
2. Select File|New. The resulting dialog box will show the same number of cases (samples) as are in the regular data matrix.
3. Set the “Data File Type” option to CCA Environmental.
4. Set the desired number of initial variables.
5. Press OK.

6. Go to Data|Edit Data menu and choose CCA Env. Data from the resulting submenu. This will open a data editor showing your new environmental data matrix, complete with sample labels. Note that when there are two files open you can have two data editors, one for each file. The commands on the Edit and Data menus apply to whichever editor is active.
7. Enter the environmental data in the same way as for regular data and save them.

The steps for loading a second data matrix for CCA are as follows:

1. Load (or create) a MVS file.
2. Choose the File|Open menu item.
3. In the resulting dialog box change the file type so it is set to the "CCA environ. files" option.
4. Now select one of the MVE files.
5. Press OK. The status window will show a second pane giving the details of this second data file. You will receive a warning if the two files opened do not have the same number of samples.

To perform a CCA with these data follow these steps:

1. Create or load the two files, as described above.
2. Choose the Analyses|Correspondence Analysis menu item.
3. Click on the "Canonical CA" option so that it is ticked.
4. Set any other options, as required (see full description of the CA dialog)
5. Press OK. The analysis will be performed and the results placed in the Results window.
6. Graphs of the results may now be produced in the same way as other results. The graph dialog for CCA results will allow you to plot either the samples, the variables (species) or both as a joint plot. By default vectors will be plotted for

the environmental variables. To leave the vectors off click the “Include environmental vectors” so that it is unticked.

7. Some variables may represent classes of a nominal variable (such as treatment regime or sampling site type; see full description of CCA for more details). These are best plotted with each class as a single point at the centroid of the samples belonging to that class, rather than as vectors. The CCA dialog has a list of all environmental variables. Tick the ones that represent nominal variables and they will be plotted as points, not vectors.
8. Press OK. The graph will be produced. It can be modified just like other scatter plots.

## Customizing the toolbar

The main MVSP window has a toolbar (also called a speedbar) across the top. By default it has buttons that allow easy access to the file opening and saving functions as well as the clipboard functions. You can add buttons for other functions as well. You can also move the toolbar to the side or bottom of the window, if you wish.

### *Moving the toolbar*

You can most easily move the toolbar by simply dragging it; click on the toolbar at some point where there are no buttons then, while holding down the mouse button, drag it to a new location. An outline will appear showing you the position where the toolbar will be docked. You can also click the right mouse button on the toolbar and choose a location from the resulting menu. Finally you can choose the Options|Preferences menu item, then specify the position in the Toolbar section of the dialog.

### *Customizing the toolbar*

The toolbar is customized through a dialog box. This can be invoked by double clicking on the toolbar at a point where there are no buttons, by right clicking on the toolbar and choosing “Customize” from the resulting menu, or by clicking the “Customize” button on the Options|Preferences dialog box.

The dialog box will show the main menu items for MVSP on the left. When you click on one of them the commands on that menu, along with associated icons, will appear on the right. To add a command to the toolbar just click on and drag a button from the dialog to the toolbar. It will be placed where you drop it. If you wish to move a button to another position on the toolbar just drag it to that new place. To remove a button drag it off of the toolbar. When you are done press “Close”.

By default the buttons have a flat appearance. When you move the mouse over a button it will rise up to indicate that it can be pressed (although if the button is disabled, because the function is not available at that time, it will remain flat). You can revert to having buttons that are three-dimensional all the time by unticking the “Flat buttons” option on the Options|Preferences dialog box.

# Chapter 2 - Analyses

This chapter gives general information about the techniques available in MVSP. For more specific information about the techniques as they are implemented in MVSP, see the appropriate sections in Chapter 5 for the individual dialog boxes used for initiating analyses.

## Principal Components Analysis

Principal components analysis (PCA) is one of the best known and earliest ordination methods, first described by Karl Pearson (1901). Mathematically, PCA consists of an eigenanalysis of a covariance or correlation matrix calculated on the original measurement data. Graphically, it can be described as a rotation of a swarm of data points in multidimensional space so that the longest axis (the axis with the greatest variance) is the first PCA axis, the second longest axis perpendicular to the first is the second PCA axis, and so forth. Thus these first few PCA axes represent the greatest amount of variation in the data set and hopefully contain some patterns of significance.

When a PCA is calculated, first the covariance or correlation matrix is calculated for the variables. The correlation matrix is used if standardization is desired; this is useful if the variables have been measured on different scales or are of different orders of magnitude. Otherwise the covariance matrix should be used. An eigenanalysis is then performed on the matrix.

There are several sets of results. First the eigenvalues are given. In PCA these equal the variance accounted for by each PCA axis. The eigenvalue for the first axis will be the largest, the second the second largest, and so on. The percentage of the total variance of each axis will also be calculated. Hopefully the first two or three axes will account for a large proportion of the variance, say

50-60% or more. In some cases the first axis might account for over 90-95% of the variance. In all but the simplest data sets this result should be looked at with skepticism. This may occur, for instance, when a few variables have very large values that are one or two orders of magnitude greater than the others. The analysis will be dominated by these few large variables. In these situations you may want to consider using a correlation matrix instead or transforming the data to logs or square roots.

Also provided will be the eigenvectors for each PCA axis. Each eigenvector is composed of values called the component loadings for that axis. Each variable in the original data matrix has a component loading associated with it in the eigenvector. These loadings may be considered a measure of the relative importance of each variable in the extracted PCA axis. The sign of the value indicates which end of the axis the variable is associated with. If, for example, variables A, C, and F have high positive loadings on the first PCA axis and variable H has a high negative loading, this means that the largest proportion of the variance in the data can be accounted for by the trends in these four variables. The different signs indicate that variable H has high values in a certain set of cases whereas A, C, and F have high values in a completely different set of cases.

The third set of results is a matrix of component scores. Again, one set of scores is provided for each PCA axis and each score corresponds to one case. These are computed by simply multiplying the component loadings by the original data. The resulting scores may be plotted on a scatterplot so that the first two PCA axes, for example, may be plotted against each other and the individual points would indicate the cases. In the example above, those cases that have high values of variables A, C, and F would be plotted at the positive end of the first axis, whereas those with variable H would be at the negative end.

## Principal Coordinates Analysis

Principal coordinates analysis (PCO) can be viewed as a more general form of PCA. Whereas in PCA the use of a covariance or correlation matrix is implicit, PCO can use a variety of different measures of distance or similarity. It then performs an eigenanalysis of the matrix, giving eigenvalues and eigenvectors. In general, the distances or similarities are measured between the cases directly, rather than the variables as in PCA, and the eigenvectors represent the scores for the cases. It thus gives a direct ordination of the cases and is useful in situations where there are more variables than cases (PCA is not recommended under this circumstance). The disadvantage is that no results are given for the variables, unlike PCA, which provides component loadings.

The main advantage of PCO is that many different kinds of similarity or distance measures can be used. For instance, if you are working with mixed data, in which some variables are measurements whereas others are binary or multi-state, Gower's general similarity coefficient can be used to combine these data. These coefficients can then be analyzed using PCO, whereas this data matrix would not be able to be analyzed by other ordination methods without recoding the data so that they are all in the same form. Alternative distance measures such as the Manhattan Metric could be analyzed as well.

PCO is restricted to analyzing distances or similarities that are metric. For a measure to be metric, it must follow several mathematical rules that will not be explained in detail here. However, these rules basically state that the distance must be able to be viewed in some sensible geometrical manner. Most importantly, the distances between a set of three points must be such that a triangle can be drawn. This means that the distance between two of the points (one side of the triangle) must be less than the combined distances that would form the other two sides of the triangle. It may seem that this must always be true, but if, for instance, a set of Pearson's correlation coefficients were treated as distances, there would be some cases where one would not be able to use them to draw a triangle.

## Distance and Similarity Measures

This procedure calculates a variety of distance and similarity measures. The distances are calculated between the rows of the data matrix. An option to transpose the data matrix is included, to allow analysis of the columns without requiring re-entry of the data. There are numerous publications that discuss different type of measures. I have primarily relied on the following in implementing the formulae used in this procedure: Prentice (1980), Sneath & Sokal (1973), Pielou (1984), Greig-Smith (1983), Gordon (1981), and Everitt (1980). You may refer to these for details about the measures provided in *MVSP*.

There are presently 23 measures available. These, and their formulae, are listed below. In these formulae, *i* and *j* represent two rows (cases) of the data matrix, *k* represents the column (variable), and therefore $x_{ik}$ would be the datum in the *k*th column of row *i*. *n* is the total number of variables.

The abbreviation in parentheses after each name is the tag that MVSP adds to the title of a .MVD file created using that measure. This allows the measure used to be identified when the file is reloaded.

***Euclidean distance (EUCLID):***

$$\mathrm{Ed}_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2}$$

***Squared Euclidean distance (SEUCLID):***

$$\mathrm{SEd}_{ij} = \sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2$$

***Standardized Euclidean distance (STEUCLID):***

$$\mathrm{StEd}_{ij} = \sqrt{\sum_{k=1}^{n}\left(\frac{x_{ik} - x_{jk}}{\mathrm{sd}_k}\right)^2}$$

where: sd$_k$ = standard deviation of all the elements of *k*

***Cosine theta (or normalized Euclidean) distance (COSINE):***

$$\mathrm{CTd}_{ij} = \sqrt{\sum_{k=1}^{n}\left(\frac{x_{ik}}{\mathrm{ss}_i} - \frac{x_{jk}}{\mathrm{ss}_j}\right)^2}$$

where: $\mathrm{ss}_x = \sqrt{\sum_{x=1}^{n} x_{xk}{}^2}$

***Manhattan metric distance (MANHAT):***

$$\mathrm{MMd}_{ij} = \sum_{k=1}^{n}\left|x_{ik} - x_{jk}\right|$$

***Canberra metric distance (CANBER):***

$$\mathrm{CMd}_{ij} = \sum_{k=1}^{n}\frac{\left|x_{ik} - x_{jk}\right|}{\left(x_{ik} + x_{jk}\right)}$$

***Chord distance (CHORD):***

$$\mathrm{Cd}_{ij} = \sqrt{\sum_{k=1}^{n}\left(\sqrt{x_{ik}} - \sqrt{x_{jk}}\right)^2}$$

***Squared chord distance (SCHORD):***

$$\mathrm{SCd}_{ij} = \sum_{k=1}^{n}\left(\sqrt{x_{ik}} - \sqrt{x_{jk}}\right)^2$$

***Chi-square distance (formula X2 of Prentice, 1980) (CHISQR):***

$$\mathrm{CSd}_{ij} = \sqrt{\sum_{k=1}^{n} \frac{\left(x_{ik} - x_{jk}\right)^2}{\sum_{l=1}^{n} x_{lk}}}$$

***Average distance (AVERAGE):***

$$\mathrm{Ad}_{ij} = \sqrt{\frac{\left(\sum_{k=1}^{n} x_{ik} - x_{jk}\right)^2}{n}}$$

***Mean character difference distance (MEANCHAR):***

$$\mathrm{MCDd}_{ij} = \frac{\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|}{n}$$

***Bray Curtis Distance (BRAYCURT):***

$$\mathrm{BCd}_{ij} = \frac{\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|}{\sum_{k=1}^{n} \left( x_{ik} + x_{jk} \right)}$$

***Pearson product moment correlation coefficient (PEARS):***

$$\mathrm{PCc}_{ij} = \frac{\sum_{k=1}^{n} \left(x_{ik} - \overline{x_i}\right)\left(x_{jk} - \overline{x_j}\right)}{\sqrt{\sum_{k=1}^{n} \left(x_{ik} - \overline{x_i}\right)^2 \sum_{k=1}^{n} \left(x_{jk} - \overline{x_j}\right)^2}}$$

### *Spearman rank order correlation coefficient (SPEAR):*

$$\mathrm{SCc}_{ij} = 1 - \frac{6\sum_{k=1}^{n}\left(R_{ik} - R_{jk}\right)^2}{n^3 - n}$$

where: $R$ = rank order of element in variable

### *Percent similarity coefficient (PERCENT):*

$$\mathrm{PSc}_{ij} = 200\frac{\sum_{k=1}^{n}\min\left(x_{ik}, x_{jk}\right)}{\sum_{k=1}^{n}\left(x_{ik} + x_{jk}\right)}$$

where: min = minimum of two values

### *Modified Morisita's coefficient (MODMORIS)*

$$\mathrm{MMc}_{ij} = \frac{2\sum_{k=1}^{n} x_{ik}x_{jk}}{\left[\sum_{k=1}^{n}\left(x_{ik}{}^2 / N_i{}^2\right) + \sum_{k=1}^{n}\left(x_{jk}{}^2 / N_j{}^2\right)\right] N_i N_j}$$

### *Gower general similarity coefficient (GOWER):*

$$\mathrm{GGSc}_{ij} = \frac{\sum_{k=1}^{n}\left(w_{ijk}s_{ijk}\right)}{\sum_{k=1}^{n} w_{ijk}}$$

where: $\mathrm{s}_{ijk} = 1 - \frac{\left|x_{ik} - x_{jk}\right|}{\mathrm{range}(k)}$ for quantitative data

= 1 for matches of binary or multistate data

= 0 for all mismatches

$w_{ijk}$ = 0 for negative matches of binary data

= 1 in all other situations

For this coefficient, the data type for each variable (column) must be declared. This is done through the first two characters of the data labels: those beginning with 'B_' are taken to be binary and those with 'M_' are multi-state; anything else is considered quantitative. For instance a variable indicating the presence or absence of sepals in a flower would have the label B_SEPAL, that indicating the color of the petals (one of four possible) would be named M_COLOUR, and petal length would be recorded in the column with the label LENGTH.

***Binary measures***

The following binary (presence/absence) coefficients are based on a table of frequency of matches and mismatches of the presence or absence of a single variable. The binary data should be entered into the data matrix as 0 (zero) and 1 (one). Any number that is not zero is also treated as a one, indicating presence.

| | *Sample j* | |
|---|---|---|
| *Sample i* | **Presence** | **Absence** |
| **Presence** | a | b |
| **Absence** | c | d |

***Sorensen's coefficient (SOREN):***

$$\mathrm{Sc}_{ij} = \frac{2a}{(2a+b+c)}$$

***Jaccard's coefficient (JACCA):***

$$\mathrm{Jc}_{ij} = \frac{a}{(a+b+c)}$$

***Simple matching coefficient (MATCH):***

$$\mathrm{SMc}_{ij} = \frac{(a+d)}{(a+b+c+d)}$$

***Yule coefficient (YULE):***

$$\mathrm{YMc}_{ij} = \frac{(ad - bc)}{(ad + bc)}$$

***Nei & Li's coefficient (NEI):***

$$\mathrm{NLc}_{ij} = \frac{2a}{(a + b) + (a + c)}$$

***Baroni-Urbani Buser (BARONI)***

$$\mathrm{BUBc}_{ij} = \frac{\left(\sqrt{ad} + a\right)}{\left(a + b + c + \sqrt{ad}\right)}$$

## Correspondence Analysis

PCA was developed with the assumption that the data are in the form of continuous measurements, such as length and width. It can be applied to other data, such as counts or presence/absence, with satisfactory results, but it is really not designed for this. Correspondence analysis (CA), on the other hand, was specifically developed to deal with data in the form of contingency tables, in which the number of different objects (for example, taxa) in each case is enumerated.

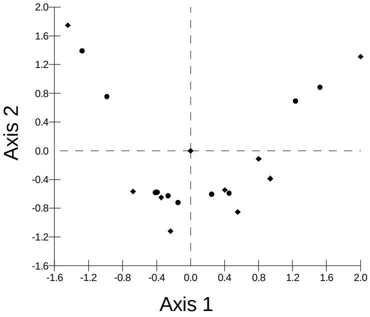
This method is known by several names, including reciprocal averaging, dual scaling, and *analyse factorielle des correspondences*. It may be calculated in a manner similar to PCA, using eigenanalysis. However, rather than performing an analysis of a covariance or correlation matrix, a similarity matrix is produced by transforming the raw data by the column and row totals of the matrix. This in effect produces a similarity matrix based on the chi-square distance, rather than the Euclidean distance implicit in the covariance matrix of PCA. The eigenanalysis is then performed on this matrix, as in PCA.

CA can also be performed through an iterative process referred to as reciprocal averaging. In this, case scores are estimated from the variable scores through a weighted averaging process, then new variable scores are estimated from the resulting case scores. This is then repeated, continuing until the two sets of scores stabilize. This gives the case and variable scores for the first axis. The process can then be repeated, with modifications to take into account the results on the first axis, to calculate the second axis scores, continuing until the desired number of axes have been extracted. Both methods of calculating CA will give the same results, although they may be scaled differently. The end product is two sets of scores, one for the cases and one for the variables. These can either be plotted on scatter graphs separately or as joint plots, as described more fully in the section on results scatter plots (Chapter 3).

In many comparative studies of ordination methods in ecology, CA is found to perform better (i.e. it more accurately summarizes the structure of the original data) than PCA, particularly when species turnover is high and some cases at one end of a gradient have very little similarity with those at the other. However, CA is particularly susceptible to dominance by unusual, outlying cases or species. For instance, when one or more cases are very different from all others in the analysis, the first few axes will often have these outliers at one end and all the other cases packed close together at the other. This problem can be rectified by removing the outlying cases. If you wish to retain these cases in the analysis, an alternative approach is to ignore the first axes that exhibit dominance by outliers and focus attention on subsequent axes.

### *Detrended correspondence analysis*

CA is also susceptible to two faults that are common to many ordination methods. The first and most prominent is what is called the arch effect or alternatively the horseshoe effect. With this effect, the points are arranged in an arched pattern along the first two axes, rather than a linear pattern as would be expected (see example below). This arch is a result of the data reduction process and represents a mathematical relationship between the first two axes, which are supposed to be independent. The effect is particularly pronounced when a long environmental gradient has been sampled, so that cases from one end are mostly or completely different from those at the other.



The second fault, which is a result of the first, is the compression of data points at the ends of the axes. Pairs of cases that are equally dissimilar will appear closer together at the ends of the axes than in the middle, thus misrepresenting the distance between these pairs. Both of these faults can be corrected with detrended correspondence analysis (DCA).

DCA corrects the arch effect in the following manner: After the first two axes are extracted with the reciprocal averaging technique the first axis is divided into several segments. The scores on the second axis for each point are then adjusted so that the mean score of the points within each segment is the same as that in other segments. This is like cutting the scatterplot into a number of vertical strips and moving each up and down until the points are in a straight line. The scores are then adjusted along the first axis so that they are more evenly spaced.

This method can often give more interpretable results, but it can also introduce distortion or instability of its own. The method can be viewed as using a hammer to smooth out distortions in a sheet of metal. It may work but it may also remove some embossed patterns that were supposed to be there. Therefore it is always a good idea to try both regular and detrended correspondence analysis on a data set and compare the results.

**Note:** The CA procedure in MVSP is not affected by the instability problems described by Oksanen and Minchin (1997; J. Vegetation Science 8:447-454). The correct algorithm is used in the procedure for non-linear rescaling when doing detrending. Also, the instability of the CA due to lax convergence criteria is not a problem so long as you have the accuracy setting on the CA dialog box set to 1E-6 or higher.

## Canonical Correspondence Analysis

All the other ordination methods in MVSP are indirect gradient analysis methods. In these, the data are subjected to some type of mathematical manipulation in order to reveal the most important trends in the data. These trends are then often compared to other data relating to the same samples to determine the relationship between the two. To give an ecological example, environmental data such as rainfall or a lake's pH would be compared to the trends seen in species data. These comparisons are often done either by eye or by performing a statistical procedure, such as regression, on the species and environmental data. Direct gradient analysis methods, on the other hand, encompass techniques for relating the species data directly to the environmental data, rather than having to go through two steps. Until recently this would mean plotting individual species against a few environmental parameters to see the patterns. However, any analysis that had more than three species or environmental variables would require several diagrams to show all the data. What is needed is an ordination method that incorporates environmental data directly into the analysis.

Canonical correspondence analysis (CCA; ter Braak, 1986,1987) is a multivariate direct gradient analysis method that has become very widely used in ecology. As the name suggests, this method is derived from correspondence analysis, but has been modified to allow environmental data to be incorporated into the analysis. It is calculated using the reciprocal averaging form of correspondence analysis. However, at each cycle of the averaging process, a multiple regression is performed of the sample scores on the environmental variables. New site scores are calculated based on this regression, and then the process is repeated, continuing until the scores stabilize. The result is that the axes of the final ordination, rather than simply reflecting the dimensions of the greatest variability in the species data, are restricted to be linear combinations of the environmental variables and the species data. In this way these two sets of data are then directly related.

The results of CCA can be presented in a diagram containing the environmental variables plotted as arrows emanating from the center of the graph, along with points for the samples and taxa. The relationships between the samples and species are as in CA; each sample point lies at the centroid of the points for species that occur in those samples. The arrows representing the environmental variables indicate the direction of maximum change of that variable across the diagram. If an arrow for pH points to the right of the diagram, this indicates that pH is increasing along a gradient from the left to the right. The length of the arrow is proportional to the rate of change, so a long pH arrow indicates a large change and indicates that change in pH is strongly correlated with the ordination axes and thus with the community variation shown by the diagram. The position of the species points in relation to the arrows (as determined by drawing perpendicular lines from the point to the arrow) indicates the environmental preference of that species.

## Cluster Analysis

Cluster analysis is a term used to describe a set of numerical techniques in which the main purpose is to divide the objects of study into discrete groups. These groups are based on the characteristics of the objects and it is hoped the clusters will have some sort of significance related to the research questions being asked.

Cluster analysis is used in many scientific disciplines and a wide variety of techniques have been developed to suit different types of approaches. The most commonly used ones are the agglomerative hierarchical methods. Hierarchical methods arrange the clusters into a hierarchy so that the relationships between the different groups are apparent. The results of this type of analysis are generally presented in a tree-like diagram called a dendrogram. The term agglomerative means that the dendrogram is produced by starting with all the objects to be clustered separate, then successively combining the most similar objects and/or clusters until all are in a single, hierarchical group.

The agglomerative clustering algorithm proceeds as follows:

1. First the similarity between each pair of cases must be calculated and placed in a matrix. There are numerous types of similarity and distance measures that can be used.
2. This matrix is then scanned to find the pair of cases with the highest similarity (or lowest distance). These will be the most similar cases and should be clustered most closely together.
3. The cluster formed by these two cases can now be considered a single object. The similarity matrix is recalculated so that all the other cases are

compared with this new group, rather than the original two cases.

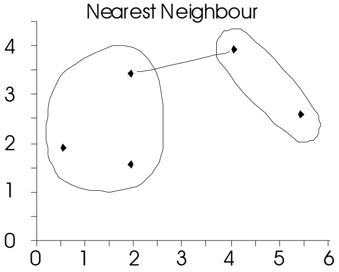
4. The modified matrix is then scanned (as in step b) to find the pair of cases or clusters that now have the highest similarity. Steps b and c are repeated until all the objects have been combined into a single group.

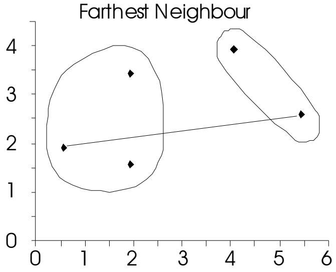
The result is a dendrogram that shows the most similar cases linked most closely together. The level of the vertical lines joining two cases or clusters indicates the level of similarity between them. It is important to note that the branching hierarchy and the level of similarity are the only important features of the dendrogram. The exact order of the cases along the vertical axis is not significant. The dendrogram can be envisaged as a mobile that allows the individual clusters to rotate around.

There are seven types of agglomerative clustering methods commonly in use. These all follow the basic algorithm outlined above, varying only in the manner in which the similarity between clusters is calculated (step c).

### *Nearest and farthest neighbor*

These two are the simplest methods. With nearest neighbor clustering, the distance between one group and another is taken as the distance between their two closest points. Farthest neighbor, on the other hand, takes the distance between the two farthest points as being that between the two groups. These methods are also known as single linkage and complete linkage respectively. The diagram below shows how the distances for these two methods would be calculate in a two dimensional example.


Nearest Neighbour

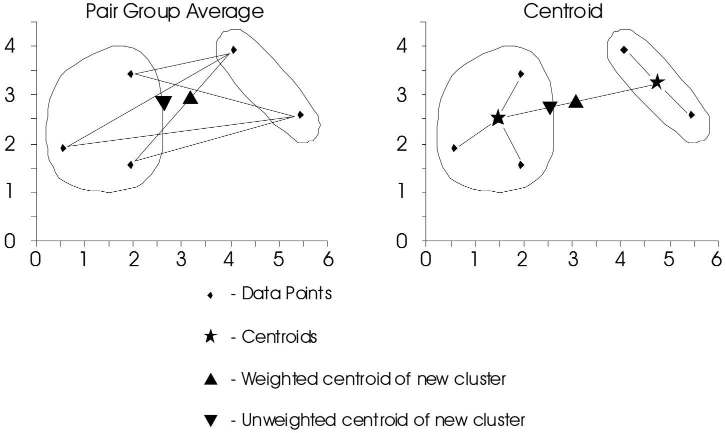

Farthest Neighbour

They may be simple, but these methods can also be viewed as distorting the data, since the distances between groups are calculated based on unusual outlying points rather than the properties of the whole cluster. Nearest neighbor clustering is also susceptible to a phenomenon called "chaining" in which there is a tendency to repeatedly add new individuals onto a single cluster rather than making several separate clusters. This gives the dendrogram a staircase-like appearance.

***Average linkage***

Average linkage techniques provide a more balanced approach to clustering. With these techniques, the distances between groups are represented as some sort of an average distance. There are two basic approaches. The distances between each pair of points in the two clusters can be measured, then the mean of these distances can be used as the distance between the clusters. This method, called the pair group average method, is illustrated below

Alternatively, the centroid of each group can be calculated and the distance between the groups is the distance between the centroids. The centroid itself can be described as the average point of the cluster. It is calculated by taking the mean value of the coordinates on each axis for all the points in the cluster. So for the example below, labeled Centroid, the centroid of the group to the left would have the coordinates 1.5 (the mean of 0.5, 2.0, and 2.0) on the X axis and 2.4 (the mean of 3.5, 2.0, and 1.7) on the Y axis.


Pair Group Average
Centroid
0
1
2
3
4
0
1
2
3
4
5
6
♦ - Data Points
★ - Centroids
▲ - Weighted centroid of new cluster
▼ - Unweighted centroid of new cluster

There are also two variants that apply to both of these methods; the calculations can be either weighted or unweighted. The unweighted methods give equal weight to each point in each cluster. The weighted methods give equal weight to each cluster instead; if one cluster has fewer points than another, those points in the smaller cluster must be given higher weighting in the calculations to make the two groups equal. The differences between these two can be visualized by plotting the centroids of the two groups on the diagrams above. With unweighted methods, the centroid is closer to the group with more points, whereas the centroid in the weighted methods is halfway between the two groups. In general, the unweighted versions are used unless the data are expected to have some clusters that are much smaller than others (e.g. if some communities have been sampled less than others).

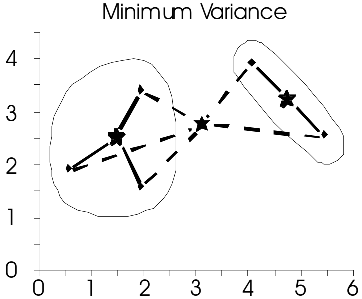
The following table lists the names of the four average linkage methods, as used in MVSP, and their classification.

| | **Pair Group Average** | **Centroid** |
|---|---|---|
| **Unweighted** | UPGMA | Centroid |
| **Weighted** | WPGMA | Median |

### *Minimum variance*

Minimum variance clustering (also called Ward's method or sum-of-squares cluster) takes a very different approach to agglomerative clustering. Instead of measuring the distances between clusters, the method focuses on determining how much variation is within each cluster. The clusters to be joined in the next round of clustering are then chosen by determining which two would give the least increase in within-cluster variation. In this way, the clusters will tend to be as distinct as possible, since the criterion for clustering is to have the least amount of variation.

When performing minimum variance clustering, first the within-group dispersion (or variance) is calculated for each cluster. This is done by first finding the centroid of each cluster. The within-group variance is calculated by taking the sum of the squared distances between the centroid and each point (the solid lines in the figure below). Next, for each pair of clusters, the centroid for all the points in the two clusters is calculated and the variance within the combined group is calculated (the dashed lines below). This is done for each pair of groups and the pair that has the lowest variance is chosen to be combined.



### *Constrained*

The types of clustering discussed so far simply join the most similar objects into clusters; the original order of the objects is not taken into account. Constrained clustering, on the other hand, forces the original order to be maintained, so that the resulting dendrogram reflects the similarity between adjacent cases. This is obviously very useful for stratigraphical studies in which the original order is of importance.

Constrained clustering can be performed as a variation of any of the above agglomerative procedures. Instead of scanning the whole distance matrix for the most similar objects, constrained clustering scans just the distances of adjacent cases. Otherwise it can be calculated as described above.

## Diversity Indices

This procedure computes three diversity indices commonly used in ecology, Simpson's, Shannon's, and Brillouin's. See Pielou (1969) and Krebs (1989) for a discussion of the use and derivation of these indices.

Two variations of Simpson's are available. The most commonly presented form, which is labeled simply "Simpson's" in the drop-down list on the dialog box, is based on the following formula:

$$D = 1 - \sum_{i=1}^{s} p_i^{\,2}$$

where: $D$ = Simpson's diversity index
$p_i$ = Proportion of species $i$ in the community
$s$ = Number of species

This version is appropriate for samples taken from an infinite population and may be applied to measures of biomass and cover as well as individual counts. However, it is a biased estimator of the true diversity index for the whole population. An unbiased version (labeled "Simpson's unbiased" in the MVSP diversity dialog) is proposed by Pielou (1969):

$$D = 1 - \sum_{i=1}^{s} \left[ \frac{n_i (n_i - 1)}{N(N-1)} \right]$$

where: $D$ = Simpson's diversity index
$n_i$ = Number of individuals of species $i$ in the sample
$N$ = Total number of individuals in the sample

This form can only be used with count data. When the number of individuals ($N$) is high the two forms will give almost identical results.

The formulae for the other two measures are as follows:

$$H' = -\sum_{i=1}^{s} p_i \log p_i$$

where: $H'$ = Shannon's diversity index

$$H = \frac{1}{N} \log\left( \frac{N!}{n_1!n_2!n_3!\ldots n_s!} \right)$$

where: $H$ = Brillouin's diversity index

As with the unbiased Simpson's, Brillouin's can only be used with counts. Shannon's can be used with any form of data.

The input data file should be set up with species as columns and samples as rows. The diversity, then, is calculated for each row. The output consists not only of the diversity index, but also the number of species and the evenness, which is defined as the diversity divided by the maximum possible diversity. See Krebs (1989) for explanations of how to calculate the maximum (and minimum) diversity.

# Chapter 3 - Graphs

## Data scatter plot

Simple scatter plots of your original data may be produced to allow you to investigate the interrelations between two or three of the variables. The scatter plot dialog box allows you to specify which variables to plot.

By default the resulting graph will be a simple 2-d or 3-d graph with triangles for the points and axes that are scaled and positioned to best fit the data. You can make extensive modifications to this graph through the graph customization dialog box (accessed through the Graph|Edit Graph menu item). Some of the more common modifications you may wish to make are as follows:

- **Modifying the shape, size and color of the data points** - This is done through the Markers page of the Edit Graph dialog box. Use the drop down boxes to select the appropriate color and shape, then use the scroll bar to adjust the size.
- **Adding labels to each point** - Go to the Labels page and make sure the “On” box in the “Data Labels” section is ticked. Note that you can also click on a data point and a temporary label will appear identifying that point.
- **Adding a title at the top of the graph** - This is done through the “Graph Title” box on the Titles page.
- **Changing the axis titles** - This is also done through the Titles page, via the Left and Bottom title boxes. Note you can adjust the orientation of the titles here as well.
- **Changing the graph’s background color** - This is done through the Background page. The “Graph Window” section lets you change the

background color for the whole graph. You may also specify a bitmap to form the background

- **Changing the color or framing of the titles and graph** - This can also be achieved through the Background page. Select the graph element you wish to modify in the "Apply To" section, then change the style and colors in the "Style" section.
- **Changing the fonts** - Do this through the Fonts page. You must first select which graph element you are changing the font of in the "Apply To" section, then change the typeface and size through the appropriate controls. "Graph Title" will change the font for the title at the top of the graph. "Other Titles" will change that for the axis titles at the left and bottom. "Labels" will change the font for the data labels as well as the numeric labels along the X and Y axes. "Legend" will change the font for the legend that displays the group name for each different symbol.
- **Modifying the position and scaling of the axes** - This is done through the Axis page. You must first select which axis you are modifying in the "Apply To" section, then use the radio buttons in the "Position" and "Scale" sections. If a "User Defined" scale is selected you can set the scale limits in the "Range" boxes. You may also adjust the tick marks, label orientation and display of grid lines here.
- **Fitting curves and other statistical lines** - This is done through the Trends page. Simply tick the boxes for the statistical lines to add and adjust their color if needed. For curve fitting you must select a method of fitting the curve and adjust the related options, if applicable.

## Results scatter plot

Scatter plots of your ordination results may be produced for any set of two or three axes. The scatter plot dialog box allows you to specify which axes to plot as well as the graph type. In addition to simple scatter plots you may also produce joint plots and biplots. The interpretation of these types of plots is discussed below.

By default the resulting graph will be a simple 2-d or 3-d graph with triangles for the points and axes that are scaled and positioned to best fit the data. You can make extensive modifications to this graph through the graph customization dialog box (accessed through the Graph|Edit Graph menu item). Some of the more common modifications you may wish to make are discussed in the previous section.

### *Joint Plots*

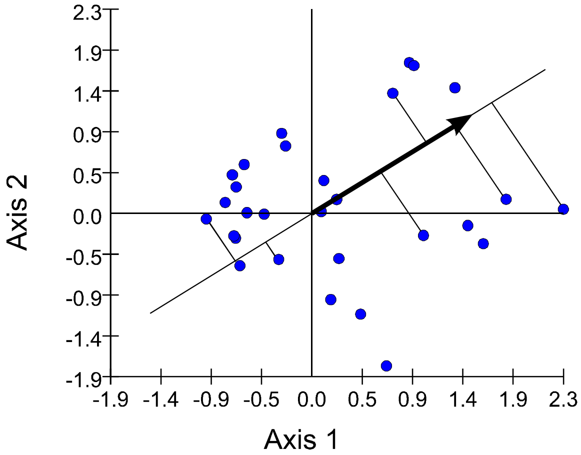
This type of plot is available for plotting the results of correspondence analysis (and canonical correspondence analysis). Since these methods ordinate the scores for the variables and cases together, the two sets of scores have equivalent scaling and can be plotted together on the same graph. MVSP lets you specify this type of graph in the scatter plot dialog box. The resulting graph will have the two sets of scores plotted together; each set will be represented by a different symbol (or color, if you so choose through the Markers page of the Edit Graph dialog).

With joint plots the investigator can easily see the relationships between the variables and cases. When the default scaling (by species) is used each case point lies at the centroid of the points for the variables associated with that case (see the section on Scaling in the Correspondence Analysis dialog box section of Chapter 5 for alternative scalings and their results). Thus the variables that characterize a case can be determined by looking at which variable points lie nearby. In doing this, though, the investigator should be aware that some of the variable points that occur in the center of the diagram may be placed there because they are unrelated to the particular ordination axes being plotted, rather than because they occur primarily in the nearby cases. These non-specific variables can easily be identified by looking at the original data matrix to see if the abundance values are similar in all cases.

### *Biplots*

Biplots have vectors superimposed over the scatter plot points. These vectors emanate from the center of the graph and represent either the variables (in PCA) or the environmental variables (in CCA). The direction of the arrow indicates the direction of maximum change for that variable and its length is proportional to the rate of change.

For example, if an arrow for pH points to the right of the diagram, this indicates that pH is increasing along a gradient from the left to the right. If it is a long arrow this indicates a large change and that change in pH is strongly correlated with the ordination axes and thus with the community variation shown by the diagram.



The position of the points in relation to the arrows (as determined by drawing perpendicular lines from the point to the arrow; see example above) indicates the relationship between each point and the variable represented by the arrow. Those that are farthest along towards the head of the arrow are those with the largest values for that variable.

Usually MVSP biplots will have a legend at the bottom giving the "Vector scaling". The position of the head of the vector is determined by the component loading in PCA or the biplot score in CCA. However, these scores often will produce arrows that either extend far beyond the cloud of points for the cases or that are very short ones at the center of the cloud. To improve the viewability of the graph the vectors are scaled so that they fit the range of the cloud of points. The vector scaling factor lets you know how much they have been changed.

### *Plotting CCA centroids*

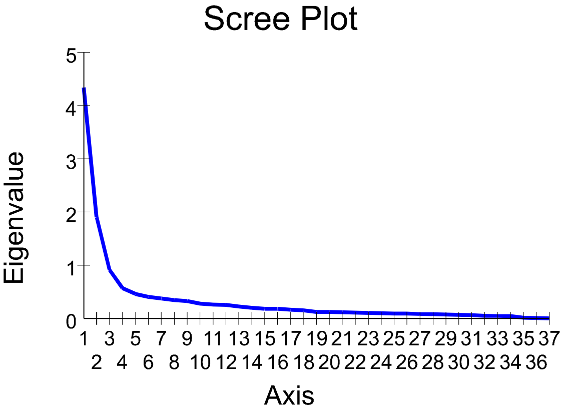
In CCA nominal variables (i.e. those that are measured as a set number of states, such as color or land use type) are best represented by a series of "dummy" variables, one for each class. For example, a land use nominal variable with three possible land uses will be coded as three separate variables. A particular sample that comes from, for instance, agricultural land will have a 1 in the data matrix for the agricultural dummy variable, and a 0 for the other two dummy variables.

These types of variables could also be represented on the biplot as vectors, but it is often considered more natural to plot each dummy variable as a point representing the centroid for that class of the nominal variable. Then the point will lie at the centroid of the samples that are from that particular type of land use.

To do this the actual biplot scores must be scaled somewhat differently (which is done automatically by MVSP). The scatter plot dialog box for CCA provides a mechanism for declaring which variables should be plotted as centroids.

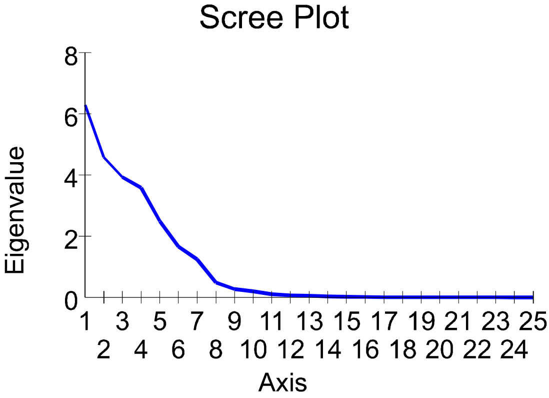
## Scree plot

The scree plot provides a means of assessing how well the variability in the data is represented by the first few axes of the ordination. It is constructed by simply plotting the eigenvalues in descending order to produce a line graph. Since the first eigenvalue is the largest, and the following ones are in descending value, this produces a line graph that slopes down to the right, as in the following example:



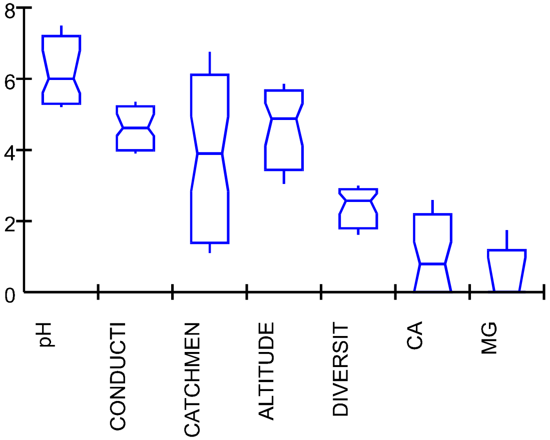
This illustrates how most of the variability is accounted for in the first few axes; in this example 64% is accounted for by axes 1-5. Most axes have much lower variability. By looking for the point where a pronounced change in the slope occurs we can decide how many axes we need to examine. In this case that point could be either axis 4 or 5.

The following example shows a scree plot in which the slope is much lower. A larger proportion of the axes should be examined in this case.



## Box and whisker plot

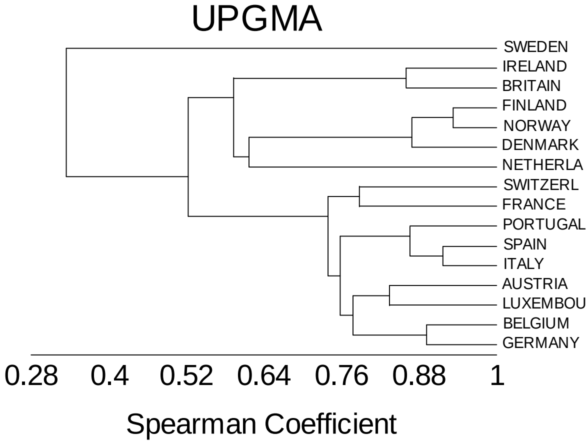
Box and whisker plots provide a graphic means of summarizing each variable in your raw data. It illustrates the spread of values about the median. Visually each variable is represented by a box with a waisted notch about the median and vertical lines (“whiskers”) extending from the top and bottom.



The notches delimit the quartiles of data. The whiskers delimit the 5th and 95th percentiles. The entire box delimits the 10th and 90th percentiles.

# Dendrogram

A dendrogram, or tree diagram, is the most common method of displaying the results of a cluster analysis. The branching pattern of the dendrogram illustrates the similarity between the various objects being clustered. The closer they are linked the more similar they are. The following is an example.



When interpreting a dendrogram it is important to remember that the only aspects of the graph that count are the branching order and the lengths of the branches. The precise order of the objects on the right side of the diagram should **not** be considered important. The dendrogram can be viewed as a mobile hanging from the ceiling; the various branches are free to be rearranged and rotated. For example, in the above graph Sweden could easily have been placed at the bottom of the dendrogram rather than the top. Likewise the group containing the countries from Switzerland to Germany could have been rearranged so that Austria was next to the Netherlands and Switzerland and France at the bottom.

# Chapter 4 - Menus

This chapter lists all of the items on the MVSP menus and gives brief descriptions of them. Most of these menu items will lead to dialog boxes. Full details of what these dialog boxes do are given in Chapter 5.

## File menu

***New***
Create a new MVSP data file. See the section about the New File dialog box (Chapter 5) for more information.

***Open***
Open an existing MVSP desktop or data file. See the section about the Open File dialog box (Chapter 5) for more information.

***Close***
Close the current data file and all associated results, graphs and notes windows. You will be prompted to save any changes that have been made to the data or desktop.

***Save Data***
Save the current data file. See the section about Saving Data (Chapter 1) for more information.

***Save Data As***
Save the current data file under a different name. See the section about the Save As (Chapter 5) dialog box for more information.

***Save Desktop***
Save the state of the current desktop, including the position and contents of all windows and any changes made to the data files. See the section about Saving Your Work in Chapter 1 for more information.

### *Save Desktop As*

Save the current desktop under a different name. See the section about the Save As (Chapter 5) dialog box for more information.

### *Import*

Import data from a spreadsheet, database or text file. See the section about the Import dialog box (Chapter 5) for more information.

### *Export*

Saves data, results, graphs or notes to disk under alternative formats so that they can be imported to other programs. See the sections about export data or graphs and results for more information.

### *Merge Files*

Merge several MVSP data files into a single one. See the section about the Merge Files dialog box (Chapter 5) for more information.

### *Print*

Print the contents of the currently active window. See the section about the Print dialog box (Chapter 5) for more information.

### *Printer Setup*

Change the selected printer and its properties. See the section about the Printer Setup dialog box (Chapter 5) for more information.

### *Page Setup*

Change the printed page margins for the currently active window. See the section about the Page Setup dialog box (Chapter 5) for more information.

### *Exit*

Exit the MVSP program. If data or results have been modified you will be prompted to save them.

### *List of file names*

A list of the most recently used data and desktop files will be listed at the bottom of the menu. A single click on any of these will load the file without displaying the Open File dialog box.

## Edit menu

### *Undo*

This reverses the last editing action performed on data or graphs. MVSP will save all of your editing actions during the current session, allowing you to undo them one by one. Note that this menu item changes its caption to give a better idea of what type of action will be undone. When the cursor is over this item the message area on the status bar will give greater detail of the action.

Note that each window maintains a separate list of editing actions that can be undone. Choosing this menu item will undo the last action performed in the currently active window. Also note that, if the results or graphs windows are closed and then later reopened you will only be able to reverse actions that have occurred since the window was reopened.

### *Redo*

If any editing actions have been redone this command will reinstate that action. For example, if some text was deleted, then restored by choosing the Undo item, the Redo item will delete it again.

### *Cut*

This menu item deletes the currently selected text or data and places it in the Windows clipboard. You can then use the Edit|Paste command to either paste it to a different place in MVSP or to place it in another program.

### *Copy*

This is like the Cut menu item, except that the selected text or data is not deleted. If the graphics window is currently active then the whole graph is copied.

### *Paste*

Paste text or data from the Windows clipboard into the currently active data editor or notepad. If the current contents of the clipboard are not text then this item will be disabled. It will also be disabled for Results and Graphics windows, since they are not editable.

### *Clear*

This menu item deletes the currently selected text or data. In the data editor it will not actually delete the cells, but will instead replace all contents with zeros. To delete whole rows or columns use the Delete Rows/Columns menu item, described below.

### *Select All*

This will automatically select all the text or data on the currently active window, making it easy to copy the entire page.

### *Delete Row/Column & Insert Row/Column*

These two menu items will produce a dialog box allowing you to delete and insert rows and columns in the current data editor.

If you have selected one or more rows or columns in the editor (by clicking on the numbered row headers on the left or the lettered column headers at the top) then these menu items will perform their insertion or deletion actions immediately, with no dialog box being shown. The captions of the menu items will change to reflect their new style of action.

If delete is chosen all columns or rows selected will be deleted. If you are inserting then the same number of columns or rows will be inserted as are selected (e.g. if you select five rows and choose Insert then five rows will be inserted).

## Data menu

### *Edit Title*

This option lets you change the title associated with the currently opened file, using the Edit Title dialog box.

### *Edit Data*

This option opens the data editor for the currently opened file. If two files are opened for CCA then you are given a choice of which to edit. The same applies to the rest of the options on this menu.

When the data editor is open this menu item will have a tick mark next to it. Choosing this item again will close the window.

Note that the rest of the options on this menu are only available when the data editor is opened.

### *Transform*

This will perform a variety of transformations (e.g. log or square root transformation) on the current file. The type of transformation and the variables to transform can be selected with the Transform dialog box.

### *Transpose*

This transposes the current data file (i.e. flips the matrix so that the columns become rows and vice versa). It has immediate action and does not produce a dialog box.

### *Convert*

This will perform a variety of conversions of the current file. The type of conversion can be selected with the Convert dialog box.

### *Add/Remove Groups*

This option will either add or remove a grouping variable, depending on whether the current data file has groups. For more information on grouping variables see the section “Working with grouped data” in Chapter 1.

***Show Group Palette***
This option controls whether the Group Palette is visible. The option is ticked when the palette is visible. Clicking the option will hide the palette, and clicking it again will make it visible again. This option can only be used when the current data file contains groups.

## Analyses menu

The items on this menu initiate analyses of the data. You will be presented with a dialog box with two or three pages, selected by tabs at the top. The first page will have the most commonly changed options, while the second will have some more specialized options that are not as often needed. Most analyses have a third page that let you select any variables and cases you wish to drop from the analysis.

For more details see the sections of Chapter 5 on the analysis dialog boxes.

## Graphs menu

Graphs of both the raw data and the results may be produced and modified with this menu.

***Scatter Plot***
This will produce 2-d and 3-d scatter plots of either the data or results (depending on which windows are open and active). See the section about the Scatter Plot dialog box (Chapter 5) for more information.

***Scree Plot***
This will produce a scree plot of the ordination results. It acts immediately, without displaying a dialog box.

***Box and Whisker Plot***
This will produce a box and whisker plot of the current data file. See the section about the Box & Whisker Plot dialog box (Chapter 5) for more information.

### *Edit Graph*

This option allows you to modify the appearance of the graph through a wide variety of options. It calls up the graph customization dialog box, containing a number of pages related to different aspects of the appearance. For hints on how to customize certain aspects see the help files sections on the appropriate graph type.

### *Zoom In*

This menu option lets you zoom in (or magnify) the currently visible graph, so that points and small labels can be viewed more clearly. It can be selected multiple times to zoom in even closer, in increments of 100% magnification (e.g. select this twice for 200% magnification, three times for 300%, etc.).

Note that the graph cannot be edited while zoomed in.

### *Zoom Out*

This will reverse the zooming process by the same increments as the zooming in option.

### *Original Size*

This option will restore the graph to its original size.

### *Adjust Vector Scaling*

PCA and CCA biplots have vectors emanating from the origin of the graph. These vectors are automatically scaled in relation to the data points so that both fit on the graph with the best amount of spread. This menu option lets you adjust the scaling of the vectors to suit your own preference.

### *Reset Defaults*

When you make changes to a graph through the Edit Graph dialog box MVSP will automatically save the settings for changed options. These settings are then used for future graphs; in this way your favorite customizations are always used. Each graph type (data scatter plot, results scatter plot, scree plot, box &

whisker and dendrogram) saves its own set of customized settings.

However, sometimes certain options will conflict. Other times changes may make a graph unviewable. Rarely, some mistaken entries can cause all future attempts to produce a graph to fail. In these cases, unless you can remember which options you changed, it is difficult to get back to a more desirable graph.

If you find yourself in this situation you can use the "Reset Defaults" menu item. This will erase all the saved customizations, so that the current and future graphs will be redrawn with the default settings, as they were when you first installed MVSP.

Note that this will not affect any graphs you have saved to desktop files.

## Options menu

### *Font*

This allows you to change the font used in the current window. See the section about the Font dialog boxes (Chapter 5) for more information.

### *Format*

This lets you change the number of decimal places to display in the current window. See the section about the Format dialog box (Chapter 5) for more information.

### *Preferences*

This lets you change several general options that affect the program operation and appearance.

See the section about the Preferences dialog box (Chapter 5) for more information.

## Window menu

***Show Notepad***

This causes the Notepad window to be displayed. When the window is open this menu item will have a tick mark next to it. Choosing this item again will close the window.

***Cascade***

This cause all the windows within the MVSP program frame to be resized to the default Windows size (with a width approximately twice the height) and arranged so they overlap in a cascading fashion.

***Tile***

This cause all the windows within the MVSP program frame to be resized so that none of them overlap and there is no empty space in the program frame.

***Arrange Icons***

If you have minimized one or more windows within the MVSP program frame, turning them into icons within the frame, this will arrange the icons along the bottom of the frame.

***Minimize All***

This will minimize all windows within the MVSP program frame to icons at the bottom of the frame.

***Close All***

This will close all the windows currently open in MVSP, leaving an empty program frame. This will also cause the data file to be closed. You will be prompted if the data or desktop need to be saved.

***List of windows***

The titles of all windows open within the MVSP program frame is listed at the end of this menu, with each preceded by a number. You can bring a window to the top and make it the active window by selecting its corresponding menu item.

## Help menu

***Index***

This menu item will open up the MVSP help file at the index or contents page.

***How to use Help***

This will display some basic information about using the Windows help system.

***Order MVSP***

This launches the MVSP Ordering Wizard. It takes you step-by-step through the process of filling out an order form for MVSP, which can be used to purchase the program after evaluation. The Wizard automatically calculates the correct price.

***About***

In the full registered version of MVSP this displays copyright information and the name of the person or organization to whom this copy of the software is registered. In the evaluation version a message is displayed showing the amount of time left in the evaluation period. It also provides a link to the ordering wizard and to a dialog box where you can enter a code to unlock the evaluation version, removing the time limitation.

# Chapter 5 - Dialog Boxes

MVSP has numerous dialog boxes for performing a variety of functions. They are listed below, arranged in the order of the menu items that invoke them. The links below will give full details about the options on the dialogs. You can also call up the help page directly from the dialog using the ? Help button on most dialog boxes.

Within dialog boxes you can use the “What’s this” help facility. Simply click on the ? button in the upper right corner. The mouse cursor will turn into a help cursor combining an arrow and question mark. Next click on the part of the dialog box in which you are interested. A small window will appear with an explanation of that item.

In the rest of this chapter the sections about each dialog box are entitled with the menu items that invoke them (e.g. the dialog for opening a file is titled File|Open, indicating that the File menu should be pulled down and the Open item selected).

## File|New

This dialog box allows you to create a new MVSP data file.

You are asked to enter the initial number of rows (cases or samples) and columns (variables). You can easily add more rows or columns later if needed, so don’t worry if you aren’t sure of the exact number needed.

The right section of the dialog box lets you specify the type of data matrix to create. Normally you would create a regular MVSP data file, which contains raw variables x cases data. Alternatively, you can create a symmetrical data matrix of distances or similarities; these can be used for input to clustering or PCO, particularly if you are importing distances or similarities calculated by another program.

A third type of data matrix is available if you currently have a regular MVSP data file open. This is a CCA environmental data file. If you choose this type a second matrix is created and opened, which can be analyzed along with the regular data matrix using CCA in order to examine relationships between the variables in the two files. When using this type of data file the number of cases in both files should be the same.

The "Include grouping variable" option lets you specify that you wish to assign group membership to each case. See the "Working with grouped data" section in Chapter 1 for more information.

When you press OK the data file will be created and the status window displayed. To add data simply choose the Data|Edit Data menu item and begin entering data.

## File|Open

This dialog box lets you load an existing data or desktop file. This is an enhanced dialog box that is similar to the standard Windows Open File dialog box but adds some useful new features.

The dialog provides a list of all MVSP files in the current directory. You can select one of these by single clicking on it, or type a name of a file into the "File name:" box. To limit the list of files to just certain types (such as desktop files or symmetrical matrix files) use the "Files of type" drop down box. Pressing OK once the filename is list in the "File name" box will open that file.

When you select a MVSP file information about that file is displayed at the bottom to make it easier to ensure you are selecting the correct file.

At the top of the dialog is a drop-down box labelled "Recent Directories". This is a list of up to 10 directories that you have recently loaded files from or saved files to. To go quickly to one of these directories, without having to navigate through the directory tree, simply click on the box and select the desired directory.

At the top is a button that looks like this: . Pressing this button will split the area where the files are listed and will show a tree-like display of your hard drives, directories and network drives, similar to that in Windows Explorer.

To open a file in a read-only state (so that it cannot be changed accidentally) make sure the "Read-only" box is ticked.

## File|Save As

This dialog box lets you save either the current data file or the entire desktop (graphs, notes and results). Which will be saved is determined by the file type set in the “Save as type” section of the dialog box. Many of the features of this dialog box are similar to the Open Dialog Box described above.

The dialog provides a list of all of the appropriate types of MVSP files in the current directory. You can select one of these by single clicking on it, or type a name of a file into the “File name:” box.

The possible file formats listed in the “Save as type” section varies depending on the currently opened file. If it is a regular MVSP data file (*.MVS) then that format will be listed along with desktop files. If the currently open file is a symmetrical file then you will instead be allowed to save it as a symmetrical file or to save the whole desktop.

If two data files are open for CCA then Save As will normally offer *.MVS as the data file type and Save As will only save the regular data file, not the environmental data file. To save the environmental file under a different name, first open the environmental data editor (Data|Edit Data|CCA Env. Data) and make sure the data editor is the active window. Then when you select File|Save As the dialog will display environmental data files (*.MVE) and allow you to save the second data matrix.

# File|Import

MVSP imports data through an import preview dialog box. This shows you a preview of how the data will be imported and allows you to change various options so that the data are imported correctly.

When you first choose the File|Import option you will first be shown a standard Open File dialog box. The "Files of type" section will list all the possible import formats. Set this to the correct format and select a file.

When you press OK MVSP will import a portion of the file. You might see one of two warning dialog boxes, one saying that non-numeric data points were found and another saying that only a certain number of data rows were imported for preview, not the whole file. You may simply click OK for both of these.

You will next be presented with the Import Preview dialog box. This has a tabbed section at the top, with various options that can be changed on different pages (described in detail below). At the bottom is a spreadsheet-like area that shows you how the data will be imported. The portions of the file that will be read as variable labels will be in the shaded row at the top, and the parts that will be the row labels (one for each case) will be in the shaded section to the left.

The actual data points will be in between. These should all be numeric data. If you got a warning that non-numeric data points were found then you will see that some cells have contents that are either blank or not numbers. This often happens when Import has not correctly located the row and column labels or it may mean that the data matrix does indeed contain non-numeric entries. When the data file is finally imported all these will be turned into zeros.

At this point you need to examine the previewed data to ensure that they are being imported correctly. If they are not then adjust the options described below to correct any problems.

### *Options*

#### COLUMN NAMES

This tells import where to find the text labels used for each column (variable). The exact options available vary depending on the file type being imported. Some file types don't offer any options here at all; for example, with database files the column names are taken from the field names.

"Guess names" (for spreadsheets only) tells import to make its best guess as to where the labels are. If this does not find the names then you can use "Names at row" to specify which row contains the names. If there are no column names in the file then use "Make up names" to create column names of the form Col1, Col2, etc.

#### GROUPING VARIABLE, COLUMN:

If the data file being imported contains a grouping variable then you can use this option to let MVSP know which column has the group names. This is done by ticking the checkbox and setting the number field at the right to the number of the column. The column will be shaded in the preview grid.

#### IMPORT AS STRINGS

Import will normally try to interpret any columns of numeric data as sensible floating point numbers, but you may find in some cases this will not represent the data correctly. Ticking "Import as Strings" may change the resulting display of the numeric data. Try ticking and unticking this option to see what effect it has on your data.

#### CREATE NEW ROW NAMES

Import will normally take the first column of entries in the imported file as row labels. If your file does not have row labels, then some of your data may be turned into labels. To avoid this, tick the "Create new row names" option to create row labels of the format Row1, Row2, etc.

### MAX. PREVIEW ROWS

When you change any option on this dialog the data will be reimported to reflect the new choices. With large files this could take a long time. To avoid this you can limit the number of rows to preview to a smaller number, such as 500 or 100.

## *Text Options*

These options are only available for text file importing.

### FIELD DELIMITER

With text files the columns of data must be separated by some character. MVSP will attempt to determine the type of character delimiting the columns, but it may not be successful. If you don't see columns of numbers in the preview box then check the setting of this option to make sure it matches your file. If you aren't sure of the format you can experiment with different options to see which works best. Be particularly aware that, if the numbers in the imported file use a comma for the decimal point rather than a period (full stop) setting the delimiter to a comma will result in the data being imported completely wrong.

### QUOTE CHARACTER

Often text files will have labels (such as for the variables and cases) surrounded by quotes. This is particularly important for labels containing spaces, so that the whole string is treated as one label. This option allows you to specify the character used for quotes around labels.

### DECIMAL POINT

This option allows you to specify the character used for the decimal point.

### *Spreadsheet Options*

**SHEET**

Most modern spreadsheet programs allow multiple pages or sheets. This option provides a list of the pages in the current spreadsheet. When you select a different page from the list the data on that page will be previewed. Note that only one page can be imported at a time.

**NAMED RANGE**

Many spreadsheet programs allow you to give a name to a certain block of cells (a range). This option will list all the named ranges in the current spreadsheet. By selecting one of these MVSP will import just the data within that range.

**RANGE**

If you have not named a range you can still specify that the import is limited to a certain block of cells by entering a range in this box. Ranges are entered using the standard spreadsheet notation, e.g. B1..K25 or B1:K25.

## File|Merge Files

This dialog box allows you to merge a number of MVSP raw data files into a single file. This allows you to easily combine data for an overall analysis.

When data files are combined variables with identical labels (ignoring differences in upper and lower case) are combined into one variable. Cases from the different files are all treated separately. You can optionally have the name of each file appended to the case labels, so that cases with identical names can be distinguished. In the following example, note that the variables A and C have been combined:

File1.mvs

```
          A      B      C      D      E
Case1     0.0    2.0    1.0    2.0    7.0
Case2     2.0    3.0    2.0    2.0    6.0
Case3     0.0    4.0    3.0    2.0    5.0
Case4     0.0    6.0    4.0    2.0    4.0
Case5     7.0    7.0    5.0    3.0    3.0
```

File2.mvs

```
          A      C      X      Y      Z
Case1     0.0    2.0    1.0    2.0    7.0
Case2     2.0    3.0    2.0    2.0    6.0
Case3     0.0    4.0    3.0    2.0    5.0
Case4     5.0    6.0    4.0    2.0    4.0
Case5     0.0    7.0    5.0    3.0    3.0
```

Merged file

```
            A     B      C      D      E      X      Y      Z
Case1 File1 0.0   2.0    1.0    2.0    7.0    0.0    0.0    0.0
Case2 File1 2.0   3.0    2.0    2.0    6.0    0.0    0.0    0.0
Case3 File1 0.0   4.0    3.0    2.0    5.0    0.0    0.0    0.0
Case4 File1 0.0   6.0    4.0    2.0    4.0    0.0    0.0    0.0
Case5 File1 7.0   7.0    5.0    3.0    3.0    0.0    0.0    0.0
Case1 File2 0.0   0.0    2.0    0.0    0.0    1.0    2.0    7.0
Case2 File2 2.0   0.0    3.0    0.0    0.0    2.0    2.0    6.0
Case3 File2 0.0   0.0    4.0    0.0    0.0    3.0    2.0    5.0
Case4 File2 5.0   0.0    6.0    0.0    0.0    4.0    2.0    4.0
Case5 File2 0.0   0.0    7.0    0.0    0.0    5.0    3.0    3.0
```

To merge files follow these steps:

1. Specify the directory containing the files (either type it in the box labeled “Directory”, or click the “Open folder” button to the right of this box and use the resulting folder browser to find the directory).
2. All the .MVS files in that directory will be shown in the “Select Files” list box. To include a file in the merged file click on the box preceding the file name so that a tick mark appears in the box.
3. To create a new MVSP file make sure the “Output File Type” is set to MVSP.

4. To append the file name to each case label tick the “Add filename to case labels” box.
5. If you wish the variables in the resulting merged file to be alphabetical then tick the “Sort Variables” checkbox. Otherwise the variables will be in the same order as the original files.
6. Press OK. The files will be read and merged and loaded into MVSP ready for analysis or further editing.

### *Producing RASC files*

You can also merge MVSP files to produce input files for the quantitative stratigraphy program RASC (Ranking and Scaling), commonly used in paleontological studies related to oil exploration. The input files consist of a dictionary listing all the stratigraphic events (the highest occurrence of a fossil species or other geological event in a geological sequence), plus a series of files, one for each well, core or geological section, listing the order in which the events occurred. Also produced is a file listing the depths of each sample.

To successfully convert MVSP files to RASC files you must ensure that:

1. The case labels consist of the numerical depth at which the sample occurred
2. The samples are in stratigraphic order, so that the one nearest the top of the stratigraphic sequence is the first row and that at the bottom is the last row
3. Absence of an event in a sample is indicated by a zero; presence of an event is indicated by any other number.
4. The labels for the same event in different files are identical (be particularly careful of misspellings in long species names).

When merging files for RASC MVSP will scan through each variable (event) looking for the highest non-zero number. The depth at which that event occurs will then be recorded in the RASC files. Note that currently MVSP can only produce files of highest occurrences.

You must enter a seven-character name for the output files in the box labeled “Output filename”. MVSP will then append the appropriate three-character extension to each of the types of files produced. Note that the filename is limited to seven characters, rather than the usual eight. This is because RASC creates names for the numerous output files by taking the input filename and adding single letters to each (e.g. if the input filename is RASCOUT then the output filenames will be RASCOUTa.out, RASCOUTb.out, etc)

You must also specify whether the depths are recorded in meters or feet.

## File|Printer Setup

This is the standard Windows Print dialog box, used in many other Windows programs. Its appearance will vary depending on the version of Windows you are using and whether you have the 16-bit or 32-bit version of MVSP. This dialog allows you to view and change the current printer and its various settings and properties.

## File|Print

This is the standard Windows Print dialog box, used in many other Windows programs. Its appearance will vary depending on the version of Windows you are using and whether you have the 16-bit or 32-bit version of MVSP. It allows you to view and change the current printer and its various settings and properties. It also allows you to specify the number of copies to print and whether to print the whole document, only certain pages or (where applicable) only the selected portion.

## File|Page Setup

This dialog lets you specify the size of the margins on the page when the current window is printed. Each MVSP window has its own margin settings. You may specify the margins in either inches or centimeters.

You may also specify that a border will be drawn around the page when printed.

## Edit|Insert/Delete Rows or Columns

The Insert and Delete dialogs are very similar and will be discussed together. These dialog boxes appear when you choose the Insert Rows/Columns or Delete Rows/Columns from the Edit menu. They only appear if the currently active data editor does not have rows or columns selected; if any are selected then the insertion or deletion occurs immediately.

The Insert dialog lets you specify whether to insert rows or columns, the number of rows/columns to insert, and whether you want them to be inserted at the cursor's current position or at the end, after the last row/column. The Delete dialog is similar, except it specifies the rows/columns to delete.

## Data|Edit Title

This dialog box allows you to enter and modify the title that is saved with each MVSP file. This title allows you to provide a longer description (up to 80 characters) of what the data are. This description is then displayed in the status window when the file is opened, as well as at the top of each results report.

If two data files are open (for CCA) you can also modify the title for the environmental data file.

## Data|Transform

The Transform option allows you to choose to have the data log or square root transformed or standardized before analysis. By default all variables in the file are transformed, but you can use the list box titled “Select Variables” to restrict transformations to certain variables. To remove a variable from the list of those to transform simply click the box preceding the variable label to clear the tick mark. The “Mark All” and “Unmark All” buttons provide rapid means to mark or unmark all variables.

Most of the procedures in MVSP assume a normal distribution of the data, but this assumption is often not met. Log or square root transforming the data can reduce the skewness of the data, resulting in a more interpretable analysis. Please note that the log transformations are performed on the values x+1, rather than x. This is done to avoid computer errors when the data value is 0, since the log of 0 is undefined, and to avoid negative results when the value is less than 1.

Standardization transforms data values so that, for each variable, the standard deviation of all cases is 1 and the mean is 0. This is particularly useful if different variables have been measured on different scales.

The logratio transformation (Aitchison, 1986) was designed specifically for compositional (percentage or proportional) data. These data are affected by closure, in which the increase of one variable necessitates the relative decrease of another, even if the absolute value of the other doesn’t change. This can cause many problems in statistical analyses. The logratio transformation eliminates the closure problem by replacing the proportions with the log of the ratio between the proportion and the geometric mean of the case. In mathematical terms, this is:

$$x'_{ij} = \log\left(\frac{x_{ij}}{g_i}\right)$$

where: $x_{ij} =$ proportion of taxon $j$ in the $i^{th}$ sample

$x'_{ij} =$ transformed value

$g_i = \sum_{k=1}^{n} \frac{\log(x_{ik})}{n} =$ geometric mean

$n =$ number of taxa in the sample

It should be emphasized that, for the logratio transformation to be calculated properly, the cases MUST be the rows of the data file. Otherwise the calculations will be meaningless.

Problems arise when some of the proportions are zeros, since taking the log of zero produces an error. This is remedied in *MVSP* by replacing them with a very small value and then readjusting all other proportions so that the total is 1.0. The replacement values are calculated using Aitchison's (1986, p.269) zero replacement formula.

## Data|Convert

This dialog box allows you to perform conversions of the data to alternative representations.

***Drop "zero" rows/columns***

The first option is to automatically drop any row or column that contains all zeros. Zero rows or columns may arise if you create a new matrix that is larger than what is required but fail to delete the extra rows or columns. They can also occur when you delete variables or cases. For example, if you have some ecological samples that contain just a single species, then deleting the variable for that species will leave the corresponding cases with all zeros.

Some analyses, such as CA, will refuse to run if there are any zero rows or columns. Others will work but the results can be obscured by these zero rows and columns. It is therefore recommended that dropping zero rows/columns be performed after data entry has been finished.

### *Range through*

The Range Through conversion is of interest primarily to geological biostratigraphers. This conversion assumes that the cases are in stratigraphic order. In many cases biostratigraphers can presume that between the first and last occurrences of a species stratigraphically (i.e. between its times of evolution and extinction) the species was in existence. The Range Through conversion applies this presumption to the data set by setting all cases between (and including) the first and last occurrence to 1, indicating its presence. All other cases are set to 0.

### *Change Scale*

The options in this section convert the data in each case to a new scale. The proportion and percentage replace each data value with its portion of the total for that case. Proportions will range from 0-1 whereas percentages range from 0-100.

The Binary option changes all non-zero data values to 1, thus converting them to presence/absence data.

The Octave option converts the data to a ten point abundance scale, roughly based on log base 2. This scale was devised for visual estimation in community ecology. It may also be used to convert fully quantitative data to a simpler scale. Much of the minor variation in abundances can be viewed as stochastic noise rather than significant trends (Gauch, 1982). By breaking the data into ten classes this minor variation is eliminated and only the major 'signal' is preserved. The conversion of data percentages to the octave scale occurs as follows:

| | |
|---|---|
| 0 | = 0 |
| >0 - 0.5% | = 1 |
| >0.5 - 1% | = 2 |
| >1 - 2% | = 3 |
| >2 - 4% | = 4 |
| >4 - 8% | = 5 |
| >8 - 16% | = 6 |
| >16 - 32% | = 7 |
| >32 - 64% | = 8 |
| >64 - 100% | = 9 |

## Analyses|Principal Components Analysis

This dialog box initiates a Principal Components Analysis of the currently loaded data. It has the following options:

### *Options Page*

**TRANSPOSE DATA**

This option transposes the data before analysis (i.e. it "flips" the data matrix so that the rows become the columns and vice versa). This is a temporary transposition and does not affect the data in memory or on disk. If you wish to permanently transpose the data use the Data|Transpose menu command.

**CENTER DATA**

Normally data in a PCA are centered around the origin of the graph before analysis. This option allows you to leave the data uncentered. An uncentered data matrix is called for when there is appreciable between-axes heterogeneity. This means that different clusters of points are associated with different axes, and have little projection on other axes. This often occurs when different groups of cases have completely different sets of values for their variables, with little overlap. See Noy-Meir (1973) and Pielou (1984) for more on this phenomenon.

**STANDARDIZE DATA**

You may choose to standardize the similarity matrix before eigenanalysis (thus creating a correlation rather than a covariance matrix). Generally a centered covariance matrix is used, but if different units of measurement are used in the data matrix, these will need to be standardized, and thus a correlation matrix should be used. Standardization may also be desired in ecological studies to reduce the effects of dominant species, so that rarer species play a greater role in the resulting configuration.

### DATA TRANSFORMATION

This option transforms the data before analysis. See the section on the Data|Transform menu item for more details about the transformations available.

This is a temporary transformation and does not affect the data in memory or on disk. If you wish to permanently transform the data use the Data|Transform menu command.

### AXES TO EXTRACT

You may specify one of a number of methods for determining the number of PCA axes to display in the results. By default all axes are displayed, but if you have a large number of variables you will have a large number of PCA axes to print out, most of which will be relatively unimportant.

You may choose to enter a specific number of axes to display by setting this option to "Enter a number" and setting the appropriate number in the "Enter number of axes" box. Alternatively, there are two rules that can be followed to display only the most important axes (those with eigenvalues above a certain value). Kaiser's rule states that the minimum eigenvalue should be the average of all eigenvalues (or 1 if the correlation matrix is used). This is often considered a good rule of thumb for determining whether a component is interpretable (Legendre & Legendre, 1983). Jolliffe (1986) proposed a modification of this rule in which the minimum eigenvalue is 0.7 times the average eigenvalue. This will usually give one or more extra components over Kaiser's rule.

### *Advanced Page*

**TRANSFORMED DATA (DISPLAY)**

This option allows you to have the transformed data printed along with the results.

**SIMILARITY MATRIX (DISPLAY)**

This option allows you to have the similarity matrix (covariance or correlation, depending on the setting of the Standardize data option) printed along with the results.

**TEXT SCATTERPLOTS**

Ticking this box causes text-based scatterplots (drawn with characters like "|", "-" and "+") of the first three axes to be produced and displayed in the Results window.

**ACCURACY**

The accuracy and speed of the eigenanalysis can be controlled by using this option. Eigenanalysis in *MVSP* is performed using either the cyclic Jacobi method or the Hill reciprocal averaging algorithm. Both of these are iterative procedures that makes repeated passes through the matrix improving the accuracy of the solution. The iterations stop when a certain level of accuracy, which is supplied by the user, is reached. Greater accuracy in the solution means that more passes must be made through the matrix, therefore the program takes longer to run.

To change the accuracy you can either drag the slider with the mouse or click to either side of it to move it by one increment. You can adjust it with the keyboard by typing Alt-A to transfer the focus to the slider bar, then using the right and left arrow keys to move the slider.

The accuracy is given in scientific notation, so that 1E-6 means 0.000001 (1.0 x $10^{-6}$). An accuracy level of 1E-7 (the default value) should be suitable for most purposes. It is not recommended that you set this below 1E-6.

### HIGHLIGHT LOADINGS GREATER THAN

This option tells MVSP to highlight (with boldface type) the component loadings greater than the set value when the results are produced. This lets you tell, at a glance, which variables are most important in the analysis on each axis, aiding in interpretation. To turn off highlighting set the value to 0.00.

## *Select Page*

This page lets you select which variables and cases to include in the analysis. By default the entire data matrix is used, but you may wish to drop some variables or cases to see the effect of their exclusion.

To drop a variable or case just click on the box next to the name so that the tick mark is cleared. You may use the buttons in the middle to rapidly mark or clear all variables or cases. Note that at least two variables and two cases must be marked before the analysis can proceed.

### AUTOMATICALLY DROP ZERO ROWS/COLUMNS

When some variables are dropped you may be left with cases that have no measurements or counts for any of the remaining variables. Likewise, dropping cases can leave variables that are not represented in any of the remaining cases. This can cause problems in some analyses, particularly correspondence analysis. This option cause the program to scan the remaining cases and variables, automatically dropping any that have all zeros.

## *About the Analysis*

MVSP performs a R-mode PCA. The component loadings are scaled to unity, so that the sum of squares of an eigenvector equals 1, and the component scores are scaled so that the sum of squares equals the eigenvalue. Q-mode PCA will generally have different scaling. Note that many packages perform Q-mode PCA, and thus their eigenvectors will be scaled to the eigenvalue, rather than unity. Performing a PCO in

MVSP using the Euclidean distance will give the same results as a Q-mode PCA.

In the R-mode analysis, similarity coefficients are calculated for the descriptors (or variables), which are the columns of the matrix and component scores are calculated for the cases or samples, which are the rows of the matrix.

## Analyses|Principal Coordinates Analysis

This dialog box initiates a Principal Coordinates Analysis of the currently loaded data. This is a generalized form of PCA. Whereas PCA implicitly uses either a covariance or correlation matrix, PCO allows you to input any matrix of metric values. PCO may be used with any of the distances calculated by *MVSP* except for the squared Euclidean distance. Of the similarity measures only Gower's is metric. PCO is calculated as a Q-mode eigenanalysis, and therefore only gives the eigenvectors, not scores. Note that a PCO of Euclidean distances will give the same results as a Q-mode PCA.

It has the following options:

### *Options Page*

The "Transpose", "Data transformation" and "Axes to extract option" are the same as those on the PCA dialog. The one new option on this page is "Similarity or distance". This is used to select the measure to use in PCO. Only the metric measures will be displayed. See the section on Distances and Similarities in Chapter 2 for the formulae and other information about these measures.

### *Advanced Page*

Most of the options on this page are the same as those on the PCA dialog. The one exception is the "Save distances to file" option. This lets you save a matrix of distances or similarities to a file, with a .MVD extension.

This file can then be loaded later for subsequent analyses, or imported to other programs. More information about these files is in the section on Working with Symmetrical Matrices in Chapter 1.

### *Select Page*

All the options on this page are the same as those on the PCA dialog.

## Analyses|Correspondence Analysis

This dialog box initiates a Correspondence Analysis (or Canonical CA) of the currently loaded data. It has the following options:

### *Options Page*

The "Transpose", "Data transformation" and "Axes to extract option" are the same as those on the PCA dialog.

#### DETREND

This option invokes the detrending procedure. It can only be used with the Hill reciprocal averaging algorithm; the setting of the Algorithm option (on the Advanced Page; see below) will be changed when this option is chosen.

Detrended correspondence analysis assumes that the actual data being analyzed are abundances of a set of variables (taxa in an ecological study) in a set of cases. Presence/absence data may also be used (entered as 0 and 1), but none of the data may be negative. It is also assumed that the cases come from a gradient in which different variables (taxa) characterize different parts of the gradient. Although it is most commonly used in ecology, this method may also be used in other fields where these assumptions hold, such as archaeology or market research.

#### CANONICAL CA

This indicates that you wish to perform a Canonical Correspondence Analysis. To perform this analysis you must have a second data matrix loaded, as described in the section on Performing Canonical Correspondence Analysis in Chapter 1. If you do not already have a secondary data matrix loaded you will be prompted to load one when you press OK.

This option can only be used with the Hill reciprocal averaging algorithm; the setting of the Algorithm option (on the Advanced Page; see below) will be changed when this option is chosen. Also, data transposition is not available with this procedure.

Note that turning on the Canonical CA option will automatically turn off detrending. Detrending of CCA results can inherently cause numerical problems in CCA and therefore is not recommended. Any arch effect that may appear in a CCA analysis is better reduced by removing superfluous environmental variables.

### SPECIES WEIGHTING

When CA is used in ecological studies it is often common to weight the species in some way, so that more emphasis is given to rarer or more common ones. When using the Jacobi algorithm (see the Algorithm option in the Advanced Page section below), the analysis can be run with a weighting of either the rare or the common species. See Orloci (1978, pp. 152-168) for details of these methods of weighting. Also, the scores can be adjusted to percentages.

When the Hill algorithm is used it is possible to have rare species downweighted before the analysis. This can be useful if you want most weight to be given to the common species, but you still want to see how the rarer taxa are affected. Those taxa that occur in fewer than 1/5 the number of cases that the most common taxon occurs in will be downweighted. The amount that the species is downweighted is related to its frequency of occurrence.

### *Advanced Page*

The "Results to display" and "Accuracy" options on this page are the same as those on the PCA dialog.

### ALGORITHM

*MVSP* normally uses the cyclic Jacobi method of calculating ordinations. This method calculates the scores for all axes simultaneously. However, the detrending process cannot be performed with this algorithm, since each axis must be detrended against the final scores of the previous axis. Also, Canonical CA cannot be used with this algorithm, since it requires that each axis be repeatedly regressed against the environmental variables as it is being extracted. Thus an alternative algorithm is used in which the solution for each axis is calculated separately. This is done using the reciprocal averaging method described by Hill (1978). The two algorithms are referred to as "Cyclic Jacobi" and "Hill" respectively.

Hill reciprocal averaging can also be used for non-detrended and non-canonical analyses as well. The algorithm only extracts the specified number of axes and is usually much faster than the eigenanalysis by the cyclic Jacobi algorithm, which must extract all axes. This is most pronounced with large data sets. However, you often need to see more than the first few axes, particularly if they do not account for much of the total variability in the data set. Examining the last axis can often be useful for identifying outliers. Also, in cases where two or more of the axes have similar eigenvalues the reciprocal averaging method may not give accurate results. If this happens a warning message will be displayed.

The actual scores produced using the two algorithms will differ, because the scaling is different, but the configuration on a plot will be the same.

### DETRENDING SEGMENTS

This option sets how many segments the axis should be divided into for the detrending process. The default value, 26, should be adequate for most analyses, but if the detrending does not seem to be as effective as it could be a larger number can be tried.

### DETRENDING CYCLES

When detrending is in force, the axes can also be rescaled so that the points at either end are not closer together than those in the middle. This rescaling is done several times and this option allows you to vary the number of times. It is generally not advisable to change this from the default of 4, however, as further rescaling may reduce the effectiveness of the ordination. Rescaling may be bypassed by entering 0 for this option.

### SCALING

When the Hill algorithm is in use (and detrending is not in force) a variety of scaling options can be used to adjust the final scores. These different methods of scaling do not affect the actual order of points along the extracted ordination axes, but they do affect the relative amount of scatter of the points between axes.

The default scaling method is "By Species", which in full means "species scores are weighted mean sample scores". This should be used in most cases and is best if your main interest is in the configuration of the species (variables). When both species and samples are plotted together the species will be plotted at the approximate centroids of the samples containing them and the distances between species will approximate their chi-square distances. If a CCA is being performed then the biplot scores (used to draw the environmental vectors on the scatterplot) will equal the correlation between that variable and the constrained CCA axis.

The "By Samples" scaling (or "sample scores are weighted mean species scores") is the opposite and is best if you are more interested in the configuration of the samples. Samples will lie at the centroids of the species they contain and the distances between samples will be chi-square distances. The "Symmetric" scaling is a compromise between these two.

The three Hill scaling options are similar to the three mentioned above except that the method of standardization is a bit different. Rather than standardizing by the eigenvalue (λ) they are standardized by λ/(1-λ). This is the method that was used in the program DECORANA, which first introduced the Hill algorithm, and earlier versions of its descendant CANOCO. These methods have the disadvantage that the positions of the species in relation to the samples and environmental vectors just indicates the relative ordering of their abundances, whereas with the first three scaling methods the approximate values of the species abundances can be inferred.

When detrending is used then there is no scaling option. The sample scores produced will be scaled to the standard deviation of the species abundance along the gradient represented by the axis. Therefore we can interpret the numbers on the axes of the scatter plot as standard deviation units. If we assume species abundance along a gradient is normally distributed, then a species will appear, rise to its highest abundance, and disappear in about 4 standard deviation units (sd). Thus if the ordination axis is relatively short (less than 3-4 sd units) then the species turnover along the gradient will be low, whereas long axes (say 12 sd units) will probably have completely different sets of species at either end. As with the "By Samples" scaling above, each sample point will lie near the centroid of it constituent species.

### *Select Page*

All the options on this page are the same as those on the PCA dialog.

## Analyses|Cluster Analysis

This dialog box initiates a hierarchical agglomerative Cluster Analysis of the currently loaded data. It has the following options:

### *Options Page*

The "Transpose" and "Data transformation" options are the same as those on the PCA dialog. Note that clustering is done of the cases. If you wish to cluster the variables you will need to transpose the data.

#### CONSTRAINED CLUSTERING

Normally the actual order of objects in the dendrogram is not important. However, if you are working with sequential data (such as in stratigraphical geological studies), a special constrained form of cluster analysis can be used (Birks & Gordon, 1985; Kovach, 1993). When this option is chosen, clustering proceeds as usual except that the objects to be fused are constrained to be adjacent in the data matrix. Therefore, the dendrogram that is produced will have the objects in the same order as the input matrix.

This type of constraint can often cause distortion in the dendrogram. In particular, reversals often occur where the distance (and therefore the branching level) between two objects is greater than that between the cluster of those two and the next object in the hierarchy. In sequences where there is a lot of variability, this can cause the dendrogram to be almost uninterpretable.

#### CLUSTERING METHOD

This allows you to choose which of the seven clustering algorithms will be used. See the full discussion of cluster analysis in Chapter 2 for details about these methods.

#### SIMILARITY OR DISTANCE

This is used to select the measure to use in clustering See the section on Distances and Similarities in Chapter

2 for the formulae and other information about these measures.

### *Advanced Page*

The "Transformed data" and "Similarity matrix" options are the same as those on the PCA dialog.

#### SAVE DISTANCES TO FILE

This lets you save a matrix of distances or similarities to a file, with a .MVD extension. This file can then be loaded later for subsequent analyses, or imported to other programs. More information about these files is in the section on Working with Symmetrical Matrices in Chapter 1.

#### CLUSTERING REPORT

Causes a report of the progress of the clustering to be printed. The average similarity or distance of the two objects or groups that have just been joined is printed out, along with their names and the number of objects in the newly fused group. If a single object is added to another cluster, the label for that object (or a numerical label corresponding to its position in the data matrix) is printed out. If a whole group is added, the node at which that group was last added to is printed out.

#### TEXT DENDROGRAM

The produces a text-based dendrogram, drawn with the characters "|", "-" and "+". This type of dendrogram has the advantage that very large dendrograms can be printed out over several pages of paper and the labels can be easily read. On graphic dendrograms with many objects the labels will be so small as to be unreadable.

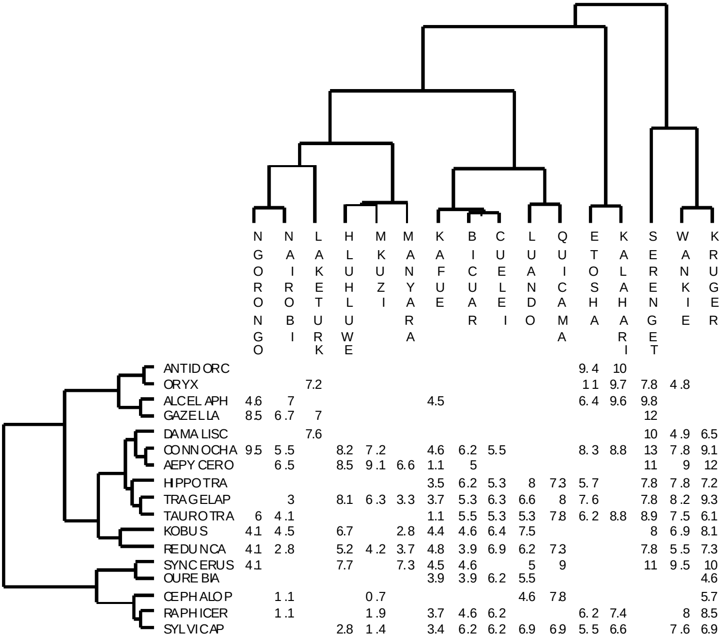
#### RANDOMIZE INPUT ORDER

The input order of the data matrix can affect the results of clustering with certain types of data sets. Changing the input order can not only change the order of objects in the dendrogram but more importantly can also cause some objects to be joined to different clusters. This is particularly possible when two or more pairs of objects have identical similarities either at the beginning or after recalculation during the clustering procedure.

Normally the clustering procedure scans through the similarity matrix sequentially looking for the next pair of objects to fuse. Choosing the “Randomize” option causes the matrix to be scanned in a random order that changes each time the procedure is run. In order to check for chaotic behavior in clustering, try running two or three clusterings of the same data matrix with this option set, then compare the dendrograms. Note that changes in the actual order of objects in the dendrogram are to be expected; a cluster diagram can be viewed as a ‘mobile’ hanging from a ceiling in which the different clusters can rotate around. It is the branching order in the dendrogram that is important and this is what should be compared when testing for chaotic behavior.

### DUAL CLUSTERING

This option causes the program to automatically do clustering of both the cases and the variables (normally just the cases are clustered). It also produces, on a separate Results window page, a copy of the original data matrix with the rows and columns sorted in the same order as the two dendrograms. You can combine this with the two dendrograms (using a separate drawing package) to produce a dual clustering diagram, as illustrated below.



| | NGORONGO | NAIROBI | LAKETURK | HLUHLUWE | MKUZI | MANYARA | KAFUE | BICUAR | CUELEI | LUANDO | QUICAMA | ETOSHA | KALAHARI | SERENGET | WANKIE | KRUGER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ANTIDORC | | | | | | | | | | | | 9.4 | 10 | | | |
| ORYX | | | 7.2 | | | | | | | | | 11 | 9.7 | 7.8 | 4.8 | |
| ALCELAPH | 4.6 | 7 | | | | | 4.5 | | | | | 6.4 | 9.6 | 9.8 | | |
| GAZELLA | 8.5 | 6.7 | 7 | | | | | | | | | | | 12 | | |
| DAMALISC | | | 7.6 | | | | | | | | | | | 10 | 4.9 | 6.5 |
| CONNOCHA | 9.5 | 5.5 | | 8.2 | 7.2 | | 4.6 | 6.2 | 5.5 | | | 8.3 | 8.8 | 13 | 7.8 | 9.1 |
| AEPYCERO | | 6.5 | | 8.5 | 9.1 | 6.6 | 1.1 | 5 | | | | | | 11 | 9 | 12 |
| HIPPOTRA | | | | | | | 3.5 | 6.2 | 5.3 | 8 | 7.3 | 5.7 | | 7.8 | 7.8 | 7.2 |
| TRAGELAP | | 3 | | 8.1 | 6.3 | 3.3 | 3.7 | 5.3 | 6.3 | 6.6 | 8 | 7.6 | | 7.8 | 8.2 | 9.3 |
| TAUROTRA | 6 | 4.1 | | | | | 1.1 | 5.5 | 5.3 | 5.3 | 7.8 | 6.2 | 8.8 | 8.9 | 7.5 | 6.1 |
| KOBUS | 4.1 | 4.5 | | 6.7 | | 2.8 | 4.4 | 4.6 | 6.4 | 7.5 | | | | 8 | 6.9 | 8.1 |
| REDUNCA | 4.1 | 2.8 | | 5.2 | 4.2 | 3.7 | 4.8 | 3.9 | 6.9 | 6.2 | 7.3 | | | 7.8 | 5.5 | 7.3 |
| SYNCERUS | 4.1 | | | 7.7 | | 7.3 | 4.5 | 4.6 | | 5 | 9 | | | 11 | 9.5 | 10 |
| OUREBIA | | | | | | | 3.9 | 3.9 | 6.2 | 5.5 | | | | | | 4.6 |
| CEPHALOP | | 1.1 | | | 0.7 | | | | | 4.6 | 7.8 | | | | | 5.7 |
| RAPHICER | | 1.1 | | | 1.9 | | 3.7 | 4.6 | 6.2 | | | 6.2 | 7.4 | | 8 | 8.5 |
| SYLVICAP | | | | 2.8 | 1.4 | | 3.4 | 6.2 | 6.2 | 6.9 | 6.9 | 5.5 | 6.6 | | 7.6 | 6.9 |

This type of diagram is useful for investigating how the data are affecting the clustering and for seeing the patterns and trends in the data themselves. Remember, though, that these are two separate analyses and are not mathematically related.

### *Select Page*

All the options on this page are the same as those on the PCA dialog.

## Analyses|Diversity Indices

This dialog box initiates calculation of a diversity index for the currently loaded data.

It has the following options:

### *Options Page*

The "Transpose" option is the same as those on the PCA dialog.

#### DIVERSITY INDEX

This lets you specify which index to calculate. See the discussion on diversity indices in Chapter 2 for more information.

#### LOG BASE

These diversity indices use logarithms in their calculation. This option allows you to specify whether to use logarithms to the base 10, 2, or e.

### *Advanced Page*

All the options on this page are the same as those on the PCA dialog.

### *Select Page*

All the options on this page are the same as those on the PCA dialog.

## Graphs|Scatter Plot

This dialog box is used to create scatter plots of either the raw data or the results of PCA, PCO and CA/CCA analyses. The exact appearance of the dialog will vary depending on the analysis.

Note that, when choosing the Graphs|Scatter Plot menu item, the dialog and graph you get depends on the active window. If you have just opened a data file and not run any analyses (or have closed the Results window) then you will only get scatter plots of the data. If analyses have been run then the page of results that is currently visible will be graphed, if possible. If the visible page is a cluster analyses or diversity calculation (for which scatter plots cannot be drawn) then the data will be plotted. Also, the data will always be plotted if the data editor is the active window.

### *Axes to plot*

All scatter plot dialogs have an “Axes to plot” section, with three drop-down boxes listing the possible elements that can be graphed on the two or three axes of the plot. For data scatter plots these will list the variables in the current file. For plots of analysis results the axes that have been extracted will be listed.

To produce a 2-d scatter plot click, in turn, on the X and Y drop down boxes to reveal the list of variables or ordination axes, then click on the ones you want on the X (horizontal) and Y (vertical) axes. If you wish to create a 3-d scatterplot change the entry in the Z drop down box from the default “None” to the appropriate setting.

### *Plot Type*

PCA, CA and CCA have a “Plot type” section that lets you specify which sets of results to plot. The first two options of all three are to do simple scatter plots of the resulting analysis scores for either the cases or variables.

CA and CCA have a third option to produce joint plots. These plots have the scores for both cases and variables plotted together, with a different symbol for each. See the section on Joint Plots in Chapter 3 for details on how to interpret these graphs.

It is not advisable to do simple joint plots of PCA scores, since the scaling of the two sets of scores is different. You can, however, produce biplots, where the cases are plotted as points and the variables plotted as vectors emanating from the origin of the graph. See the section on Biplots in Chapter 3 for details on how to interpret these graphs. Note that biplots can only be produced with 2-d scatter plots.

### *Don’t display short vectors*

Many of the vectors for biplots may be very short, since only a few variables might have high scores on the two particular axes plotted. This can create a lot of congestion around the center of the plot, as several vectors and their labels overwrite each other. This can be avoided by selecting this option, which will ensure that only vectors that are longer than 1/10 the total length of either axis will be displayed.

### *Include environmental vectors*

Biplots can also be produced from CCA results. In this case, however, the vectors represent the environmental variables. Enabling this option (which is the default) will produce biplots from CCA results.

### *Plot as centroids*

In CCA the various classes of a nominal variable are best plotted as centroids rather than vectors (see the section on plotting centroids in Chapter 3 for more details). You may declare which variables you wish to plot as centroids by selecting them from the "Plot as centroids" list box. Simply click on the box preceding each variable label, which will insert a tick mark indicating that item is selected.

## Graphs|Box and Whisker Plot

This dialog produces a box and whisker plot of the current data file. For details on interpreting these see the section on box and whisker plots in Chapter 3.

The main part of the dialog gives a list of all the available variables. By default all are selected, meaning that boxes will be produced for each and every variable. If you wish to only plot certain variables you may do this by clicking the box preceding those you do not want to plot, clearing the tick mark from the box. You can rapidly mark and unmark all variables with the "Mark all" and "Unmark all" buttons.

## Graphs|Edit Graph

This is a multi-page dialog box that allows you to adjust many characteristics of the graph. For a full description of the box click on the Help button on this dialog to go to the graph customization help file, then click on the picture of the dialog for more details.

Please note that MVSP does not implement some of the pages displayed in that help file.

## Options|Font

This dialog box lets you change the font used for text in the currently active window. For the data editor, notepad and results windows the standard Windows font dialog is displayed. You can change the font, font style and size for each window. Note that this applies to the whole window, not any selected text. Also, changing the color will have no effect.

If the graph window is active then the Fonts page of the Edit Graph dialog box will be displayed. It allows you to change the fonts of four different elements of the graph. To change each element first make sure that element is selected in the "Apply To" section, then change the "Typeface" and "Size" settings.

The four graph elements listed on the dialog box are:

Graph Title - The title at the top of the graph

Other Titles - These are the titles at the bottom, left and right of the graph. On scatter and scree plots, they form the titles of the X and Y axes. On dendrograms it is the title below the scale at the bottom.

Labels - These are the labels on the scales of the scatter, scree and box and whisker plots. For dendrograms the Labels settings apply to the labels for each object in the dendrogram.

Legend - On MVSP scatterplots this changes the font for the legend identifying the group associated with each symbol type. On dendrograms it changes the font for the scale below the dendrogram.

Above the Font Size control is a box labeled "SmartScale". When this is ticked (for any of the graph elements above) it prevents the font from being set to a size too large for the available space on the graph. This is particularly important for the labels on dendrograms with a large number of cases. With SmartScale on the font could be made very small to match the space allocated for each branch of the dendrogram. Turning this off will allow the labels to be made bigger. However, in this case the labels will then overlap. If you encounter problems with reading labels on large dendrograms then please see the "Text dendrogram" option described under "Analyses|Cluster Analysis" above. You can also use the zoom feature (see section on the Graphs menu in Chapter 4) to zoom in on the labels if they are difficult to see.

## Options|Format

This dialog box lets you change the number of decimal places to display in the current window (data editor and results window only). Simply change the number in the box to the desired number of places.

## Options|Preferences

This lets you change several options that affect how the program appears and operates.

***Automatically save and restore desktop***

MVSP allows you to save the current position and contents of all the windows visible within the program frame (referred to as the desktop). You can save the desktop only when needed through File|Save Data. However, You can also use the "Automatically save and restore desktop" option to ensure that the desktop is saved each time you exit MVSP. The same desktop will then be reloaded when you next start MVSP.

***Create backup data files***

Selecting this option causes MVSP to create a copy of a data file, containing the previously saved version, just before you save the current changes. This lets you recover the previous version if necessary. Backup files of regular MVSP data files will have the extension *.~MS, environmental data backup files will have *.~ME, and symmetrical files will have *.~MD.

### *Show logo on background*

By default the MVSP logo is embossed on the background of the window frame. If you wish to turn this off you may deselect this option.

### *Entries on recent file list*

Sets the maximum number of recently used files that will be listed at the end of the File menu. Setting this to 0 will remove the recently used file list.

### *Toolbar*

The radio buttons in this section allow you to specify the position of the toolbar (You can also drag the toolbar to the desired position with the mouse).

Pressing the “Customize” button will open a dialog box allowing you to add, remove and reposition buttons on the toolbar. See the section on “Customizing the Toolbar” in Chapter 1 for more details.

The “Flat buttons” options lets you decide whether you want regular 3-d buttons or flat ones that rise up when the mouse passes over them.

### *Display grid*

Both the data and results are displayed in spreadsheet-like grids. By default the data editor displays a grid of lines separating the cells, while the Results window does not show these grid lines. You can use the two options in the section to change the visibility of the grid lines.

### *Symmetrical matrix file type*

This specifies the format to be used for saving symmetrical matrices to disk. See the section on symmetrical files in Chapter 1 for the differences between the formats.

# Chapter 6 - Data File Structure

Normally any manipulation of MVSP data files is done through the program itself. However, there may be times when you wish to create or modify the files directly, using other programs. This section describes the structure of MVSP files.

Data files for MVSP are plain text (i.e. ASCII) files. This means that they should consist only of letters or numbers, spaces, and most other symbols represented on the keyboard. If you are modifying MVSP files with a word processor then you must make sure to save them as text files, not the word processor's native format.

All the elements (labels, data, other information) of the data file must be separated by spaces or tabs. The data files consist of three parts, the header, the column labels and the data.

### *Data File Header*

The first line of the data file should be a header line, which will give the program some information about the data, such as the number of rows and columns. It should look something like this:

```
*MVSP3 10 15
```

This header line should begin with an asterisk (*) in the first column of the first line of the file, followed by MVSP3. This identifies the file as a MVSP version 3 data file. If this is not found MVSP will not read the file. Earlier versions of MVSP had headers like this:

```
*L 10 15
```

These earlier files also had the data oriented differently (more on this later). MVSP 3 will still read these older files and will convert them to the version 3 format when saved. You can save files in this older format through File|Export.

The two numbers in the header are the number of rows and columns (not including the labels) in the data matrix. The above example has 10 rows and 15 columns.

A title may also be added to your data file on the header line, so that you know what these data represent. Here's an example

```
*MVSP3 10 15 Test data file for MVSP
```

This title will be listed to the screen and placed on the output when that file is selected. It must be separated from the other elements of the header by at least one space, and it cannot be more than 79 characters long.

### *Column Labels*

The second line contains labels for the columns, which are the variables. The number of labels must match the number of columns listed in the header. Each label may be up to 60 characters long. If a label contains a space (e.g. Quercus alba) it must be contained within double quotes ("Quercus alba") so that it is not treated as two labels. The labels may continue on to the third and subsequent lines of the file, if required.

If the data matrix contains a grouping variable (see section entitled "Working with grouped data" in Chapter 1) then the file will have an extra column before the data columns. This column will have the label "Groups$". If MVSP finds this as the first column label in this section then it will expect to read the group labels described below.

Note that this extra grouping column is **not** included in the number of columns in the data file header.

### *Data Matrix*

After the column labels comes the data matrix itself, starting on a new line. First comes a label for the row (or case), identifying that case. These row labels are required, as are the column labels. If a grouping variable is included, then the next entry in the row will be the group name. The data points come next, separated by at least one space or tab. The data should be all numeric and must consist of the numerals 0-9, plus a full stop (".") for the decimal point and/or a "-" sign if required. Data may also contain the letter E for use in entering data in scientific notation. For example 1.32E3 is the scientific notation for 1.32 x $10^3$ (1320.00).

The data for one row can continue on to the next line, but each new row must begin on a new line. Again, the number of data points in a row must match the number declared in the header. If it doesn't MVSP will attempt to read the label for the next row as a data point and will most likely fail.

Here is a complete example data file:

```
*MVSP3 5 10 Test data set for MVSP
COL1 COL2 COL3 COL4 COL5 COL6 COL7 COL8 COL9 COL10
ROW1 23  2  4 53  6 45  2  3 67  5
ROW2 10  2  4 34  1  4  3 10 20  3
ROW3  2 34  0  1 35 12  1 90 10  9
ROW4 98 12 10  4 10  9 10  5 20 31
ROW5  1  7  9 11 75  7  5 21  0 10
```

### *Symmetrical Files*

Cluster analysis and PCO create an intermediate symmetrical data matrix of similarities or distances, which are then further analyzed with clustering or ordination. You may save this symmetrical matrix to a file, perhaps for analysis with another program or to allow for future MVSP analyses without having to recalculate the distances or similarities. You can also take a symmetrical matrix created by another program and modify it to MVSP's symmetrical data format, so that you can analyze that matrix with MVSP.

These symmetrical files have a file name extension of .MVD. Here is an example:

```
*MVSP3 10 SIM Test data set for MVSP
COL1 COL2 COL3 COL4 COL5
COL6 COL7 COL8 COL9 COL10
 1.00
-0.15  1.00
 0.36 -0.05  1.00
 0.20 -0.97  0.05  1.00
-0.60  0.67  0.15 -0.60  1.00
 0.30  0.21 -0.31 -0.00  0.10  1.00
 0.30 -0.05  0.97  0.00  0.10 -0.50  1.00
-0.80  0.62 -0.41 -0.70  0.60 -0.30 -0.30  1.00
 0.82 -0.55 -0.03  0.62 -0.82  0.41 -0.10 -0.87  1.00
 0.10  0.67  0.67 -0.60  0.70  0.10  0.60  0.10 -0.41
1.00
```

As with raw data files, the "*MVSP3" is a tag that indicates this is a MVSP file. This is followed by a space, then a single number that represents the number of both rows and columns in the symmetrical matrix. The next part indicates whether this is a distance (DIS) or similarity (SIM) matrix. The rest of the first line is a title for the file and can be any text you wish.

With MVD files produced by MVSP, a short tag is added to the title, preceded by a hyphen. This tag identifies the distance or similarity measure used to create that file, thus allowing MVSP to know the type of measure when it reloads the file. If this tag is not present then MVSP will display "Unknown measure" on the output and graphs. The tags for each measure are given in the section on Distances and Similarities in Chapter 3. The following header would appear on a matrix of Euclidean distances:

```
*MVSP3 10 SIM Test data set for MVSP - EUCLID
```

Next come the labels for each column (and row). The number of these must match the number you declared in the first line. They do not need to be all on the second line; they may continue onto the third and further lines, as shown above.

As with rectangular data files, the presence of a grouping variable is indicated by a "Groups$" column label as the first label. Again, this column is not included in the number of data columns listed in the header. The group names are listed, starting on a new line, below the column labels. There should be one group label for each column/row in the matrix.

This is followed by the rows of data. The data may be in one of five formats (as described in the section on symmetrical matrices in Chapter 1). The example above is a lower half matrix with a diagonal. When reading a file MVSP will attempt to determine the type of matrix used, but may need to ask you to specify the type. The type of file produced when saving symmetrical matrices is controlled by an option on the Options|Preferences dialog box.

# Chapter 7 - Technical Support

Registered users of MVSP are eligible for free technical support on the use of the program for 90 days after purchase. This is primarily meant to answer questions about installation of the program and getting started in using it. Most questions will be answered in the manual and help files, so look at those first before contacting us. In particular, be sure to read the Getting Started section thoroughly.

After 90 days we will generally try to help with more technical questions, but we reserve the right to refer users to the MVSP or Windows manuals for more basic queries.

Please note that free technical support only applies to questions on the actual use of the program and its features; it does not extend to help on the statistical techniques themselves or interpretation of the results. The user should refer to the literature about multivariate analysis. However, Kovach Computing Services does offer data analysis consulting services at reasonable rates. Please contact us for details.

*Bug reports*

If you encounter a problem that you think may be a bug in MVSP, please contact us so that we can try to track down the cause. If the bug involved an error message from MVSP you can press the “Report” button on the dialog box, which will provide a space for you to type in details of the problem, then print out a bug report form.

Please send as much detail about the problem as possible: what you were doing at the time, a detailed description of what happened when the problem occurred, details of your system configuration and other software that was running on your machine at the time, and a copy of the data file, if possible.

*Upgrades*
Existing users of MVSP will be eligible for special prices on new versions of the program. If your name is on our registered user list you will automatically receive notification and an upgrade voucher when new versions are released. For the quickest means of getting news about upgrades join our e-mail mailing list.

To join the KCS-ANNOUNCE mailing list send an e-mail message to:

listserver@kovcomp.com

with the following text as the subject or the first line of the body of the mail message:

subscribe kcs-announce

## Hints for MVSP 2 users

If you've previously used MVSP 2 for DOS you will find that, even though the user interface has been thoroughly modernized, most of the options and analyses are similar. Here are a few hints to help you make the adjustment to version 3.

The orientation of the data matrices is now different. Previously the variables were rows; now they are the columns. This is more in line with other statistical programs. When you open an old MVSP file you will be warned that the file format has changed and the file will be converted. If you then save the file it will be saved in the version 3 file format. Note that the version 3 format is very similar to version 2, with two differences. 1) the orientation of the data is now different 2) the first line begins with *MVSP3, not *L as before. If you need to save data in the old MVSP 2 format you can do this through the File|Export command.

The options on the MVSP 3 analyses dialogs generally reflect those on the version 2 menus as much as possible. The means of setting the options differs but in most cases they perform the same function. One subtle but important difference is that the Minimum Eigenvalue option is now Axes to Extract. When you enter a number for this you are entering the number of axes to display; in the DOS version you entered a minimum eigenvalue and all axes with eigenvalues greater than this were displayed.

Clustering and PCO are now one step processes; there is no need to separately calculate the distance or similarity measure.

## Frequently Asked Questions

This section gives the answers to some common questions about MVSP. A more up to date list can be found at http://www.kovcomp.com/support/.

**Q.** Is there a way to label each point on a scatterplot with the corresponding label from the data file?

**A.** Yes.

1. Choose the Graphs|Edit Graph menu option and go to the Data page.
2. In the Data Labels section tick the box marked “On”.
3. Press OK.

Note that you can also display the label for individual points (one at a time) by clicking on the point with the mouse.

**Q.** The points on my PCA scatter plot are all clustered around the center of the second axis, rather than spread out. Why?

**A.** Ordination results are, by default, plotted with the same X and Y scales, so that the dispersion of points along the two axes is comparable. As a result, the graphs are best displayed and printed as square graphs, rather than rectangular. In your plot the amount of variance accounted for by the second axis is probably much less than for the first one, so the distribution of the points reflects this.

If you wish to use different scales on the two axes, so that the data points are spread out as much as possible, you can do this through the Axis page on the Edit Graph dialog box:

1. Choose the Graphs|Edit Graph menu option and go to the Axis page.
2. In the Apply to Axis box select the X option.
3. In the Scale section choose either zero or variable origin (depending on whether you want to force the scale to include zero).
4. Repeat step 2, choosing Y Primary instead, then repeat step 3.
5. Press OK.

**Q.** When I compare the CA/CCA results of MVSP with those of CANOCO (using the same scaling) many of the numbers are different beyond the third or fourth decimal place. Why is this?

**A.** Very early versions of CANOCO (v. 3.12 and earlier) used a fairly lax criterion for determining when to stop calculations. This has been pointed out in an article by Oksanen & Minchin (1997, J. Vegetation Sci. 8, 447-454). If you set the accuracy of a MVSP analysis (on the Advanced tab of the CA dialog) to a low level such as 1E-5, you will find that the results are identical to CANOCO. Later DOS versions of CANOCO, and the current Windows version, do not have this problem.

**Q.** Ever since I recently experimented with some of the graph customization options all attempts at creating a graph have failed with various error messages. This only happens with one graph type. What's gone wrong and how can I fix it?

**A.** When you make changes to a graph the settings are saved, so that you can create your own style of graphs and have it automatically used for future graphs.

However, certain combinations of graph customization settings seem to create problems for drawing new graphs. This also only seems to happen on certain computers. The easiest way to fix this is the use the Graphs|Reset Defaults menu option. This will erase all graph customizations and redraw all current graphs with the default settings. This will not affect any graphs you have saved to desktop files.

**Q.** I have manually created a MVSP data file from the output of another program, but I keep getting errors trying to load the file. What's wrong?

**A.** MVSP files are plain text files that can be edited with any text editor. This allows for easy manipulation of data with other programs. However, it is also easy to subtly change the structure of a file so that it can't be read properly anymore.

The structure of MVSP data files is described in Chapter 6. If you are manually creating or modifying MVSP files then you will want to understand the basic principles as to how they are structured.

The most common reasons for errors in reading MVSP data files are as follows:

1. The number of rows and columns specified in the header do not match the actual number of data rows and columns. The labels **do not** count as a row or column, these numbers should reflect the data only.
2. Either the row or column labels have been omitted. Both must be present, with the column labels preceding the entire data set and the row labels at the start of each row of data.
3. One or more of the labels or data values has a space in the middle, so that MVSP reads it as two labels or values. In MVSP 3, any labels with spaces must be surrounded by double quote marks ("). Spaces are not allowed at all in earlier versions of MVSP.
4. All zero values must be included, not left blank.

5. The data values must be separated by spaces or tabs, not commas or other characters.
6. For symmetrical matrices the matrix must follow one of the five possible formats described in the section on symmetrical matrices. MVSP attempts to detect the type of matrix being read; if it fails then you will be asked the type of matrix. If you choose the wrong type (or MVSP detects the type wrongly) then you may get errors reading the file.

If checking and correcting these aspects does not produce readable files then please e-mail one of the files to **support@kovcomp.com**. We will take a look at it to see what is wrong.

## References to Analyses

Aitchison, J., 1986. *The Statistical Analysis of Compositional Data.* Chapman and Hall, London.

Bayer, U., 1985. *Lecture notes in earth sciences. 2 Pattern recognition problems in geology and palaeontology*. Springer-Verlag.

Beals, E.W., 1984. Bray-Curtis ordination: An effective strategy for analysis of multivariate ecological data. *Advances in Ecological Research,* 14:1-55.

Birks, H.J.B., 1987. Multivariate analysis in geology and geochemistry: An introduction. *Chemometrics and Intelligent Laboratory Systems*, 2:15-28.

Birks, H.J.B. & Gordon, A.D., 1985. *Numerical Methods in Quaternary Pollen Analysis.* Academic Press, London.

Cooke, D., Craven, A.H., & Clarke, G.M., 1982. *Basic Statistical Computing*. Edward Arnold (Publishers) Ltd., London.

Davis, J.C., 1986. *Statistics and Data Analysis in Geology, 2nd Edition.* John Wiley & Sons, New York.

Duigan, C. A. & Kovach, W.L., 1991. A study of the distribution and ecology of littoral freshwater chydorid (Crustacea, Cladocera) communities in Ireland using multivariate analyses. *Journal of Biogeography,* 18:267-280.

Everitt, B., 1980. *Cluster Analysis. 2nd Edition.* Gower Publishing Co., Hampshire, 136 pp.

Felsenstein, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution,* 39:783-791.

Gauch, H.G. Jr., 1982. *Multivariate Analysis in Community Ecology.* Cambridge University Press, New York.

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

Greig-Smith, P., 1983. *Quantitative Plant Ecology.* University of California Press, Berkeley.

Hill, M.O., 1973. Reciprocal averaging: An eigenvector method of ordination. *Journal of Ecology,* 61:237-249.

Hill, M.O. 1979a. TWINSPAN - A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Cornell University, Ithaca, New York.

Hill, M.O. 1979b. DECORANA - A FORTRAN program for Detrended Correspondence Analysis and Reciprocal Averaging. Cornell University, Ithaca, New York.

Hill, M.O., & Gauch, H.G. Jr., 1980. Detrended correspondence analysis: An improved ordination technique. *Vegetatio,* 42:47-58.

Jolicoeur, P., & Mosimann, J.E., 1960. Size and shape variation in the Painted Turtle. A principal component analysis. *Growth,* 24:339-354.

Jolliffe, I.T., 1986. *Principal Components Analysis.* Springer-Verlag, New York.

Jongman, R.G.H., ter Braak C.J.F. and van Tongeren, D.F.R. 1987. *Data analysis in community and landscape ecology*. Pudoc, Wageningen.

Kovach, W. L., 1988. Multivariate methods of analyzing paleoecological data. *In:* W. A. DiMichele & S. L. Wing (eds.), Methods and applications of plant paleoecology. *The Paleontological Society Special Publication,* 3:72-104.

Kovach, W.L., 1988. Quantitative palaeoecology of megaspores and other dispersed plant remains from the Cenomanian of Kansas, USA. *Cretaceous Research,* 9:265-283.

Kovach, W.L., 1989. Comparisons of multivariate analytical techniques for use in pre-Quaternary plant paleoecology. *Review of Palaeobotany and Palynology,* 60:255-282.

Kovach, W.L., 1993. Multivariate techniques for biostratigraphical correlation. *Journal of the Geological Society, London,* 150:697-705.

Kovach, W.L. & Batten, D.J., 1994. Association of palynomorphs and palynodebris with depositional environments: quantitative approaches. *In:* Traverse, A. (ed.), *Sedimentation of Organic Particles.* Cambridge University Press. p.391-407.

Kovach, W.L., 1995. Multivariate data analysis. *In:* Maddy, D. and Brew, J. (eds.), *Statistical modelling of Quaternary science data.* Quaternary Research Association, Cambridge.

Krebs, C.J., 1989. *Ecological Methodology*. Harper Collins Inc., New York. 654pp.

Lance, G.N. & Williams, W.T., 1966. A generalized sorting strategy for computer classifications. *Nature,* 212:218.

Legendre, L., & Legendre, P., 1983. *Numerical Ecology.* Elsevier Scientific Publishing Company, New York.

Lespérance, P.J., 1990. Cluster analysis of previously described communities from the Ludlow of the Welsh Borderland. *Palaeontology,* 33:209-224.

Manly, B.F.J., 1994. *Multivariate statistical methods. A primer, 2nd edition*. Chapman & Hall, London.

Noy-Meir, I., 1973. Data transformations in ecological ordination. I. Some advantages of non-centering. *Journal of Ecology*, 61:329-341.

Orloci, L., 1978. *Multivariate Analysis in Vegetation Research, 2nd edition.* W. Junk, Boston.

Pielou, E.C., 1969. *An Introduction to Mathematical Ecology.* Wiley-Interscience, New York.

Pielou, E.C., 1984. *The Interpretation of Ecological Data.* Wiley-Interscience, New York.

Prentice, I.C., 1980. Multidimensional scaling as a research tool in Quaternary palynology: A review of theory and methods. *Review of Palaeobotany & Palynology,* 31:71-104.

Sneath, D.H., & Sokal, R.R., 1973. *Numerical Taxonomy.* W.H. Freeman & Co., San Francisco.

ter Braak, C.J.F., 1986. Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology,* 67:1167-1179.

ter Braak, C.J.F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology,* 67:1167-1179.

ter Braak, C.J.F. 1987. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 64:69-77.

# Other software from Kovach Computing Services

Brief descriptions of software provided by Kovach Computing Services are given below. More detailed information, including prices, may be found on our web pages (http://www.kovcomp.com/ or by e-mail from info@kovcomp.com). To order use the order form on our web site or contact us at the address at the start of this manual.

### *Accent Composer - A Windows keyboard utility*

Accent Composer simplifies the process of entering accented characters and other symbols into any Windows 95 or NT program. Instead of typing codes on the numeric keypad or pasting from the Character Map utility, Accent Composer lets you create accented letters with easy-to-remember two character sequences. For example, to produce an À, just press the hotkey, followed by ` and A. You can customize the hotkey as well as any of the two character sequences. Requires Windows 95/98 or Windows NT 3.51 or higher and 0.7Mb hard disk space.

### *SIMSTAT for Windows - General Purpose Statistical Package*

SIMSTAT, the popular, inexpensive and easy statistical package, is now available for Microsoft Windows(tm). This new version gives you all the power of the original SIMSTAT, but with a greatly enhanced user interface, including an innovative Results Notebook for compiling your statistical results alongside notes and annotations. SIMSTAT (written by Normand Peladeau, Provalis Research, Montreal) is an easy and powerful statistical program. It performs a wide variety of statistical analyses, including summary statistics, crosstabulation, frequencies analysis, breakdown analysis, multiple responses analysis, time series analysis, oneway analysis of variance, paired and independent sample t-tests, Pearson correlation matrix, covariance and cross product deviation, linear and nonlinear regression analysis, multiple regression analysis, GLM Anova/Ancova, single-case experimental design analysis, reliability analysis, sensitivity analysis, various nonparametric analysis, nonparametric association matrix, and bootstrap analysis. SIMSTAT will plot hi-resolution graphs and has a powerful batch command language for automating analyses. It reads data from dBase, ASCII, and SPSS files.

SIMSTAT can be integrated with a number of other statistical programs, allowing them to work from within the SIMSTAT menu structure. All results are placed in the SIMSTAT results window. SIMSTAT can be interfaced with both the Windows and the DOS version of **MVSP**. Other addins are: **WordStat** - a content analysis module specifically designed to analyze textual information; **Statitem** - a module for scale development and testing that allows one to perform classical item analysis on multiple-choice item questionnaires. Keep an eye on our web pages at http://www.kovcomp.com/ for new add-ins as well as special prices for purchasing packages of several add-ins..

***XLStat - Statistical add-in for Excel spreadsheets***

Now you can beef up the statistical capabilities of your Microsoft Excel(tm) spreadsheets with the new XLStat addin (written by Thierry Fahmy, Paris). This lets you perform multivariate techniques such as discriminant analysis, correspondence analysis and clustering, as well as a range of regression techniques, goodness of fit tests and tabular sorting, along with factor analysis, generalized non-linear fitting and exact tests. Installing XLStat gives you a new top-level menu with 25 main types of analyses. Simply choose one of these, highlight the spreadsheet range containing your data and set a few options. The results of the analysis will be placed in your spreadsheet, along with any appropriate graphs.

## Other software from Kovach Computing Services

### *Oriana - Orientation Analysis for Windows*

Oriana for Windows calculates the special forms of sample and inter-sample statistics required for circular data. It also graphs your data in a variety of ways, allowing you to easily demonstrate patterns. Oriana calculates the circular mean, length of the mean vector, circular standard deviation and standard error, 95% and 99% confidence limits, and Rayleigh's test of uniformity for each sample in your data file. Pairs of samples can be compared with Watson's F-test for two circular means. The overall distributions of two samples can be compared with Chi-squared tests. The data for each sample can be summarized with rose diagrams or circular histograms as well as linear histograms. The individual observations can be shown in raw data plots. Uniformity plots allow you to assess whether the data depart from a uniform distribution.

### *Data Desk*

Data Desk is a fast, easy-to-use data analysis package that has been helping people understand their data since 1986. Data Desk provides interactive tools for data analysis and display based on the concepts and philosophy of Exploratory Data Analysis. Data Desk implements many traditional statistics techniques suitable for data from planned experiments and sample surveys. However, the program's true strength is its powerful tools for data exploration. These tools simplify intuitive examination of your data. No special training in statistics is needed for these insightful graphic displays. When you explore your data with Data Desk, you will find patterns and relationships. But you will also bring to light the elements that don't fit — often the most important discovery you can make about your data. Putting Exploratory Data Analysis to work for you means displaying your data in many related ways. It means computations fast enough to try out several alternative analyses in the time that you might have expected to spend on a single analysis. It means linking all these views of your data to get a deeper understanding of the patterns, relationships, and exceptions in your data.

### *Wa-Tor - Population dynamics simulation*

Wa-Tor for Windows is population ecology simulation. You pit hungry sharks against tasty fish and see who comes out on top. You may get fluctuating populations that will exist for hundreds of generations. You control the initial number of sharks and fishes, their breeding rates, and shark starvation time. The dynamics of the populations can be watched on the graphs as well as on the Wa-Tor world map.

### *Data Analysis Consulting*

Do you have a data analysis problem but don't have the time to do it properly or would rather have an expert do it? Then contact Kovach Computing Services. We provide data analysis services with particular emphasis on environmental, ecological, geological and paleontological studies. Services include publication quality graphics and full reports describing the results and providing comments on their robustness.

# Index